DOCUMENT RESUME

ED 456 173	UD 034 279

TITLE	The Use of Tests as Part of High-Stakes Decision-Making for Students: A Resource Guide for Educators and Policy-Makers.
INSTITUTION	Office for Civil Rights (ED), Washington, DC.
PUB DATE	2000-12-00
NOTE	99p.; For "Standards for Educational and Psychological Testing," see ED 436 591. Primary drafters were David Berkowitz, Barbara Wolkowitz, Rebecca Fitch, and Rebecca Kopriva.
AVAILABLE FROM	For full text: http://www.ed.gov/offices/OCR.
PUB TYPE	Guides - Non-Classroom (055)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Decision Making; Disabilities; Educational Policy; Elementary Secondary Education; Evaluation Methods; Federal Legislation; *High Stakes Tests; Inclusive Schools; Limited English Speaking; Standardized Tests; *Student Evaluation; Test Bias; Test Reliability

ABSTRACT

This guide helps educators and policymakers in developing and implementing policies that involve the use of standardized tests as part of decision making that has high stakes consequences for students. The guide applies to standardized tests that are addressed in the "Standards for Educational and Psychological Testing" (Joint Standards, 1999). Chapter 1 provides information about the leading professionally recognized test measurement principles. Chapter 2 includes the legal frameworks that have guided federal courts and OCR when addressing the use of tests that have high stakes consequences for students. It discusses the federal constitutional, statutory, and regulatory nondiscrimination principles that apply to the use of tests for high stakes purposes. This document does not establish any new legal or test measurement principles. Furthermore, the test measurement principles described in Chapter 1 are not legal principles. However, the use of tests in educationally appropriate ways, consistent with the principles described in Chapter 1, can help minimize the risk of noncompliance with the federal nondiscrimination laws discussed in Chapter 2. Five appendixes present: a glossary of legal terms; a glossary of test measurement terms; accommodations used by states; a compendium of federal statutes and regulations; and resources and references. (Contains 27 resources.) (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 456 173

The Use of Tests as Part of High-Stakes Decision-Making for Students:

A Resource Guide for Educators and Policy-Makers

U.S. Department of Education
Office for Civil Rights

UD034279

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Richard W. Riley
Secretary
U.S. Department of Education

Norma V. Cantú
Assistant Secretary
Office for Civil Rights
U.S. Department of Education

December 2000

Permission to reprint this publication is not necessary. However, if the resource guide is reprinted, please cite it as the source and retain the credits to the original authors or originators of any of the documents contained in the Appendices. For questions about reprinting material in the Appendices, contact the author or originator of the document. The full text of the resource guide is available at the U.S. Department of Education's Office for Civil Rights' (OCR's) web page, www.ed.gov/offices/OCR. Individuals with disabilities may obtain this document in an alternate format (such as Braille, large print, audiotape, or computer diskette) on request. For more information, please contact OCR by telephone at 800-421-3481 or by e-mail at OCR@ed.gov. Individuals who use a telecommunications device for the deaf (TDD) may call OCR's TDD number at 877-521-2172.

This resource guide has been developed by OCR in an effort to assemble the best information regarding test measurement standards, legal principles, and resources to help educators and policy-makers ensure that uses of tests as a part of decision-making that has high-stakes consequences for students are educationally sound and legally appropriate. The resource guide is intended to reflect existing test measurement and legal principles. The resource guide is not intended to and does not add to, or subtract from, any otherwise applicable federal requirements. This publication supersedes any earlier drafts, notes, or other preparatory versions of this document.

UNITED STATES DEPARTMENT OF EDUCATION
OFFICE FOR CIVIL RIGHTS
THE ASSISTANT SECRETARY

December 2000

Dear Colleague:

Adherence to good test use practices in education is a shared goal of government officials, policy-makers, educators, parents, and students. In an era of school reforms that place increasing emphasis on measures of accountability, such as the use of tests as part of decision-making that has high-stakes consequences for students,[*] the need to provide practical information about good testing practices is well documented. In January 1999, the National Research Council (NRC) observed that we, in the education community, should work to better disseminate information related to good testing practices with a focus on the standards of testing professionals and the relevant legal principles that, together, "reflect many common concerns."

Sound educational policies and federal nondiscrimination laws can work together to promote educational excellence for all students and ensure that educational practices do not — intentionally or otherwise — unfairly deny educational opportunities to students based upon their race, national origin, sex, or disability. In short, federal civil rights laws affirm good test use practices. Thus, an understanding of the measurement principles related to the use of tests for high-stakes purposes is an essential foundation to better understanding the federal legal standards that are significantly informed by those measurement principles.

In order to further the goal of accurate and fair judgments in high-stakes decision-making that involves the use of tests, we are pleased to provide you with this copy of *The Use of Tests as Part of High-Stakes Decision-Making for Students: A Resource Guide for Educators and Policy-Makers*. This guide provides important information about the professional standards relating to the use of tests for high-stakes purposes, the relevant federal laws that apply to such practices, and references that can help shape educationally sound and legally appropriate practices.

---

[*] As explained throughout the guide, the primary focus is the use of standardized tests or assessments (referred to in the guide as tests) used to make decisions with important consequences for individual students. Examples of high-stakes decisions include: student placement in gifted and talented programs or in programs serving students with limited English proficiency; determinations of disability and eligibility to receive special education services; student promotion from one grade level to another; graduation from high school and diploma awards; and admission decisions and scholarship awards. The guide does not address teacher-created tests that are used for individual classroom purposes.

BEST COPY AVAILABLE

There are few simple or definitive answers to questions about the use of tests for high-stakes purposes. Tests are a means to an end and, as such, can be understood only in the context in which they are used. The education context — in which the relationship (and attendant obligations) of the educator to the student is frequently more complex than that between employer and employee — shows time and again that any decision regarding the legality of a use of a test for high-stakes purposes under federal nondiscrimination laws cannot be made without regard to the educational interests and judgments upon which the test use is premised.

Background

Throughout the 1990s, national, state, and local education leaders focused on raising education standards and establishing strategies to promote accountability in education. In fact, the promotion of challenging learning standards for all students — coupled with assessment systems that monitor progress and hold schools accountable — has been the centerpiece of the education policy agenda of the federal government as well as many states.

At the same time, the use of tests as part of high-stakes decision-making for students is on the rise. For example, the number of states using tests as a condition for high school graduation is increasing, with a majority of states projected to use tests as conditions for graduation by 2003 and several states now using tests as conditions for grade promotion.

Recently, more and more educators and policy-makers have requested advice and technical assistance from the U.S. Department of Education regarding test use in the context of standard-based reforms. The Department's Office for Civil Rights (OCR) is also addressing testing issues in a more extensive array of complaints of discrimination being filed with our office, most of them in a K-12 setting with implications for high-standards learning. OCR has responsibility for enforcing Title VI of the Civil Rights Act of 1964, Title IX of the Education Amendments of 1972, Section 504 of the Rehabilitation Act of 1973, and Title II of the Americans with Disabilities Act of 1990. These statutes prohibit discrimination on the basis of race, color, national origin, sex, and disability by educational institutions that receive federal funds.

In a similar vein, institutions in the post-secondary community in recent years have engaged in a thoughtful dialogue and analysis regarding merit in admissions and the appropriate use of tests as part of the process for making high-stakes admissions decisions. In some states, the use of tests in connection with admissions decisions has been an important element in public post-secondary education reform.

These trends highlight the salience of two recent conclusions of the NRC's Board on Testing and Assessment. The NRC observed that many policy-makers and educators are unaware of the test measurement standards that should inform testing policies and practices. These standards include the *Standards for Educational and Psychological Testing* (*Joint*

Standards), prepared by a joint committee of the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). The NRC also concluded that it "is essential that educators and policy-makers alike be aware of both the letter of the laws and their implications for test takers and test users." [National Research Council, *High Stakes: Testing for Tracking, Promotion and Graduation*, p.68 (Heubert and Hauser, eds., 1999).]

The Resource Guide

Toward this end, OCR has prepared this guide in an effort to assemble the best information regarding test measurement standards, legal principles, and resources to help educators and policy-makers frame strategies and programs that promote learning to high standards in ways consistent with federal nondiscrimination laws. Our goal is to inform decisions related to the use of tests as part of decision-making that has high-stakes consequences for students, such as when they move from grade to grade or graduate from high school. Just as we know that good test use practices can advance high standards for learning and equal opportunity, we know that educationally inappropriate uses of tests do not. If we want this generation of test-taking students (and their teachers and schools) to meet high standards, then we should insist that the tests they take meet high standards. When tests are used in ways that profoundly shape the lives of students, they must also be used in ways that accurately reflect educational standards and that do not deny opportunities or benefits to students based on their race, national origin (including limited English proficiency), sex, or disability.

The guide is organized to provide practical guidance related to the test measurement principles and applicable federal laws that guide the use of tests as part of decision-making that has high-stakes consequences for students. The Introduction to the guide provides a broad, conceptual overview of relevant principles so that those who are not familiar with test measurement principles or applicable federal laws can better understand the kinds of issues that relate to the use of tests in many contexts. Chapter One of the guide provides a detailed discussion of the test measurement principles that provide a foundation for making well-informed decisions related to the use of tests for high-stakes purposes. The *Joint Standards*, which has been approved by the APA, AERA, and NCME, is discussed in detail in this chapter. Adherence to relevant professional standards can help reduce the risk of legal liability when schools are using assessments for high-stakes purposes. Chapter Two provides an overview of the existing legal principles that have guided federal courts and OCR when analyzing claims of race, national origin, sex, and disability discrimination related to the use of tests for high-stakes purposes. These principles, as applied by the courts and OCR, underscore the importance of adhering to educationally sound testing practices. The Appendix includes a Glossary of Legal Terms, a Glossary of Test Measurement Terms, a list of Accommodations Used by States, a Compendium of Federal Statutes and Regulations, and a Resources and References section.

Central Principles

There are several central principles reflected in the text of this guide.

First, the goals of promoting high educational standards and ensuring nondiscrimination are complementary objectives. The ultimate question regarding the use of tests for high-stakes purposes, as a matter of federal nondiscrimination law and sound educational policy, centers on educational sufficiency: Is the test appropriate for the purposes used? That is, are the inferences derived from test scores, and the high-stakes decisions based on those inferences, valid, reliable, and fair for all students? In applying civil rights laws to education cases, federal courts recognize the importance of providing appropriate deference to the educational judgments of educators and policy-makers. These inquiries are not an effort to lower academic standards or alter core education objectives integral to academic admissions or other educational decisions. Rather, these inquires focus the educator and policy-maker on ensuring that uses of tests with high-stakes consequences for students are educationally sound and legally appropriate.

Second, when tests, including large-scale standardized tests, are used in valid, reliable, and educationally appropriate ways, their use is not inconsistent with federal nondiscrimination laws. Importantly, tests can help indicate inequalities in the kinds of educational opportunities students are receiving, and, in turn, may stimulate efforts to ensure that all students have equal opportunity to achieve high standards. When tests accurately indicate performance gaps, it is important to focus on the quality of educational opportunities afforded to under-performing students. The key question in the context of standards-based reforms and the use of tests as measures of student accountability is: Have all students been provided quality instruction, sufficient resources, and the kind of learning environment that would foster success?

Third, a test score disparity among groups of students does not alone constitute discrimination under federal law. The guarantee under federal law is for equal opportunity, not equal results. Test results indicating that groups of students perform differently should be a cause for further inquiry and examination, with a focus upon the relevant educational programs and testing practices at issue. The legal nondiscrimination standard regarding neutral practices (referred to by the courts as the "disparate impact" standard) provides that if the education decisions based upon test scores reflect significant disparities based on race, national origin, sex, or disability in the kinds of educational benefits afforded to students, then questions about the education practices at issue (including testing practices) should be thoroughly examined to ensure that they are in fact nondiscriminatory and educationally sound.

In short, the goal of the federal legal standards is to help promote accurate and fair decisions that have real consequences for students, not to water down academic standards or deter educators from establishing and applying sensible and rigorous standards. In fact, properly understood, the legal standards are an aid to meaningful education reform — by helping to ensure that instruction and assessments are aligned and structured to promote the high-level skills and knowledge that rigorous standards seek for all children.

Finally, while this guide focuses on the use of tests, similar principles apply to the overall decision-making process used to make high-stakes decisions for students. In fact, the NRC, APA, AERA, NCME, and others caution against making high-stakes decisions based on a single test score. "Other relevant information should be taken into account if it will enhance the overall validity of the decision." [*Joint Standards*, p.146 (1999).]

Conclusion

Recognizing the responsibility that educators and policy-makers must shoulder in making the promise of high-standards learning a reality, U.S. Secretary of Education Richard Riley in his commemoration of the 45th anniversary of the *Brown v. Board of Education* decision said, "A quality education must be considered a key civil right for the twenty-first century." This is the driving force behind OCR's continuing effort to provide assistance to policy-makers and educators as we continue to enforce federal laws that prohibit discrimination against students. Rather than creating false and polarizing "win-lose" choices on this all-important set of issues, we need to, as Secretary Riley noted, "search for common ground" — ground, that is, in this case, expansive.

We have worked with literally dozens of groups and individuals, including educators, parents, teachers, business leaders, policy-makers, test publishers, individual members of Congress, and others, to solicit input and advice regarding the scope, framing, and kinds of resources to include in this guide, and we are grateful for their time and assistance. The first draft of the testing guide was released in April 1999 and was the subject of substantial comments leading to extensive revisions. The second draft was released in December 1999 and once again received substantial comments. That draft also was independently reviewed by the NRC's Board on Testing and Assessment, which held a hearing earlier this year to discuss the draft guide and issued a letter report in June 2000 commenting on the draft. We are grateful for the NRC's tireless efforts. The third draft was released for public comment in July 2000, this time with notice of availability in the *Federal Register*. OCR has made numerous changes throughout the guide in response to comments seeking to clarify, make more accurate, or expand key sections. It is important to keep in mind that the guide is not designed to answer all questions related to the use of tests when making high-stakes decisions for students. However, working together with our education partners, we believe that we are providing a useful resource that will serve the education community as it addresses the very complex and important questions that stem from the institution of high standards and accountability systems designed to promote the best schools in the world.

Very truly yours,

Norma V. Cantú

Acknowledgements

This resource guide was developed by the U.S. Department of Education, in consultation with numerous stakeholders. The time and commitment of all those who provided comments and input are gratefully acknowledged. In particular, we want to recognize the primary drafters of this document: David Berkowitz, Barbara Wolkowitz, Rebecca Fitch, and Rebecca Kopriva (Consultant).

We also want to thank others from the U.S. Department of Education's Office for Civil Rights (OCR) and Office of the General Counsel who assisted in the development of the guide, including Scott Palmer, Arthur Coleman, Jeanette Lim, Susan Bowers, Cathy Lewis, Steve Beckner, Arthur Besner, Connie Butler, Doreen Dennis, Lilian Dorka, Marsha Douglas, Lisa Dyson, Joan Ford, Ann Hoogstraten, Jerry Kravitz, Jennifer Mueller, Jan Pottker, Barbra Shannon, Kimberly Stedman, Elizabeth Thornton, Rebekah Tosado, Judith Winston, Steven Winnick, Susan Craig, Karl Lahring, Lisa Battlia Anthony, Adina Kole, and Suzanne Sheridan. Additionally, we are grateful to the efforts of individuals, especially within OCR, who were responsible for developing earlier drafts of the document. Finally, we want to recognize the efforts of other persons within the U.S. Department of Education, the U.S. Department of Justice's Civil Rights Division, and the National Academy of Sciences' Board on Testing and Assessment, who reviewed drafts of this document and provided valuable guidance.

Contents

INTRODUCTION: An Overview of the Resource Guide 1

CHAPTER 1: Test Measurement Principles 21

CHAPTER 2: Legal Principles 49

APPENDIX A: Glossary of Legal Terms 69

APPENDIX B: Glossary of Test Measurement Terms 73

APPENDIX C: Accommodations Used by States 81

APPENDIX D: Compendium of Federal Statutes and Regulations 85

APPENDIX E: Resources and References 89

11

INTRODUCTION: An Overview of the Resource Guide

I. Introduction

When decisions are made affecting students' educational opportunities and benefits, it is important that they be made accurately and fairly. When tests are used in making educational decisions for individual students, it is important that they accurately measure students' abilities, knowledge, skills, or needs, and that they do so in ways that do not discriminate in violation of federal law on the basis of students' race, national origin, sex, or disability. The U.S. Department of Education's Office for Civil Rights (OCR)[1] has developed this resource guide in order to provide educators and policy-makers with a useful, practical tool to assist in their development and implementation of policies that involve the use of tests as part of decision-making that has high-stakes consequences for students.

> **When tests are used in ways that meet relevant psychometric, legal, and educational standards, students' scores provide important information that, combined with information from other sources, can lead to decisions that promote student learning and equality of opportunity. ... When test use is inappropriate, especially in making high-stakes decisions about individuals, it can undermine the quality of education and equality of opportunity. ... This lends special urgency to the requirement that test use with high-stakes consequences for individual students be appropriate and fair.**
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, p. 4 (Jay P. Heubert & Robert M. Hauser eds., 1999).

Chapter One of this guide provides information about professionally recognized test measurement principles. Chapter Two provides the legal frameworks that have guided federal courts and OCR when addressing the use of tests that have high-stakes consequences for students. This document does not establish any new legal or test measurement principles. Furthermore, the test measurement principles described in Chapter One are not legal principles. However, the use of tests in educationally appropriate ways — consistent with the principles described in Chapter One — can help minimize the risk of noncompliance with the federal nondiscrimination laws discussed in Chapter Two.

---

[1] OCR enforces laws that prohibit discrimination on the basis of race, national origin, sex, disability, and age by educational institutions that receive federal funds. The laws enforced by OCR are: 1) Title VI of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000d *et seq.* (2000) (Title VI), which prohibits discrimination on the basis of race, color, or national origin; 2) Title IX of the Education Amendments of 1972, 20 U.S.C. §§ 1681 *et seq.* (1999) (Title IX), which prohibits discrimination on the basis of sex; 3) Section 504 of the Rehabilitation Act of 1973, 29 U.S.C. §§ 794 *et seq.* (1999) (Section 504), which prohibits discrimination on the basis of disability; 4) the Age Discrimination Act of 1975, 42 U.S.C. §§ 6101 *et seq.* (1995 & Supp. 1999) (as amended), which prohibits age discrimination; and 5) Title II of the Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12134 *et seq.* (1995 & Supp. 1999) (Title II), which prohibits discrimination on the basis of disability by public entities, whether or not they receive federal financial assistance.

BEST COPY AVAILABLE

The guide also includes a collection of resources related to the test measurement and nondiscrimination principles discussed in the guide — all in an effort to help policy-makers and educators ensure that decisions that have high-stakes consequences for students are made accurately and fairly.

Recently, education stakeholders at all levels have approached OCR requesting advice and technical assistance in a variety of test-use contexts, particularly as states and districts use tests as part of their standards-based reforms. Also, OCR is increasingly addressing testing issues in a broader and more extensive array of complaints of discrimination that have been filed. These developments confirm the need to provide a useful resource that captures legal and test measurement principles and resources to assist educators and policy-makers.

High-stakes decisions in this guide refer to decisions with important consequences for individual students, such as placement in special programs, promotion, graduation, and admissions decisions.

As used in this resource guide, "high-stakes decisions" refer to decisions with important consequences for individual students. Education entities, including state agencies, local education agencies, and individual education institutions, make a variety of decisions affecting individual students during the course of their academic careers, beginning in elementary school and extending through the post-secondary school years. Examples of high-stakes decisions affecting students include: student placement in gifted and talented programs or in programs serving students with limited-English proficiency; determinations of disability and eligibility to receive special education services; student promotion from one grade level to another; graduation from high school and diploma awards; and admissions decisions and scholarship awards.[2]

This guide is intended to apply to standardized tests that are used as part of decision-making that has high-stakes consequences for individual students and that are addressed in the *Standards for Educational and Psychological Testing* (*Joint Standards, 1999*).[3] The *Joint Standards,* viewed as the primary technical authority on educational test measurement issues, was prepared by a joint committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education – the three leading organizations in the area of educational test measurement. The *Joint Standards* was developed and revised by these three organizations through a process that involved the participation of hundreds of testing

---

[2] The purpose of this guide is to address tests that are used in making high-stakes decisions for individual students. In addition to using tests for high-stakes purposes for individual students, states and school districts are also using tests to hold schools and districts accountable for student performance. Although the use of tests for this purpose is not the focus of the guide, we have provided some useful background information about relevant principles and federal statutory requirements.

[3] American Educational Research Association, American Psychological Association & National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (1999) (hereinafter *Joint Standards*).

BEST COPY AVAILABLE

professionals and thousands of pages of written comments from both professionals and the public. The current edition of the *Joint Standards* reflects the experience gained from many years of wide use of previous versions of the *Joint Standards* in the testing community.

The *Joint Standards*, which is discussed in more detail below, applies to standardized measures generally recognized as tests, and also may be applied usefully to a broad range of systemwide standardized assessment procedures.[4] For the sake of simplicity, this guide will refer to tests, regardless of the type of label that might otherwise be applied to them. The guide does not address teacher-created tests that are used for individual classroom purposes.

States and school districts are also using assessment systems for the purpose of promoting school and district accountability.[5] For example, under Title I of the Elementary and Secondary Education Act, states are required to develop content standards, performance standards, and assessment systems that measure the progress that schools and districts are making in educating students to the standards established by the state. The Title I statute explicitly requires that assessments be valid and reliable for their intended purpose and be consistent with relevant, nationally recognized technical and professional standards.[6] If educators and policy-makers consider using the same test for school or district accountability purposes and for individual student high-stakes purposes, they need to ensure that the test score inferences are valid and reliable for each particular use for which the test is being considered.[7]

> **Is it ever appropriate to test [elementary or secondary] students on material they have not been taught? Yes, if the test is used to find out whether the schools are doing their job. But if that same test is used to hold students "accountable" for the failure of the schools, most testing professionals would find such use inappropriate. It is not the test itself that is the culprit in the latter case; results from a test that is valid for one purpose can be used improperly for other purposes.**
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion and Graduation*, p. 21 (Jay P. Heubert & Robert M. Hauser eds., 1999).

---

[4] The *Joint Standards* notes that its applicability to an evaluation device or method is not altered by the label used (e.g., test, assessment scale, inventory). A more complete discussion about the instruments covered by the *Joint Standards* can be found in the introduction section of that document. *Joint Standards, supra* note 3, at pp. 3-4.

[5] The Goals 2000: Educate America Act supports state efforts to develop clear and rigorous standards for what every child should know and be able to do, and supports comprehensive state and districtwide planning and implementation of school improvement efforts focused on improving student achievement to those standards. *See* 20 U.S.C. §§ 5801 *et seq.* (1994). Largely through state awards that are distributed on a competitive basis to local school districts, Goals 2000 promotes education reform in every state and thousands of districts and schools.

[6] 20 U.S.C. § 6311(b)(3)(C).

[7] For example, if an assessment yields low scores because there is a major gap between the skills and knowledge being assessed and what is being taught, this does not undermine the validity of the assessment for purposes of program evaluation and accountability – indeed the purpose of the assessment may be to detect such gaps. In contrast, the existence of such a gap may raise serious concerns about the appropriateness of the use of the assessment for promotion and graduation decisions where students are being held accountable for what they purportedly have been taught.

BEST COPY AVAILABLE

While this guide focuses on the use of tests, similar principles apply to the overall process used to make high-stakes decisions for students. Indeed, the *Joint Standards* states that, in educational settings, a high-stakes decision "should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision."[8] As explained in the *Joint Standards*, "When interpreting and using scores about individuals or groups of students, considerations of relevant collateral information can enhance the validity of the interpretation, by providing corroborating evidence or evidence that helps explain student performance."[9] The *Joint Standards* also notes that "as the stakes of testing increase for individual students, the importance of considering additional evidence to document the validity of score interpretations and the fairness in testing increases accordingly. The validity of individual interpretations can be enhanced by taking into account other relevant information about individual students before making important decisions. It is important to consider the soundness and relevance of any collateral information or evidence used in conjunction with test scores for making educational decisions."[10] Used appropriately, tests can provide important information about a student's knowledge to help improve educational opportunity and achievement. However, as said by the National Research Council's (NRC's) Board on Testing and Assessment, "no single test score can be considered a definitive measure of a student's knowledge."[11]

Policy-makers and the education community need to ensure that the operation of the entire high-stakes decision-making process does not result in the discriminatory denial of educational opportunities or benefits to students.[12] Educators should carefully monitor inputs into the high-stakes decision-making process and outcomes over time so that potential discrimination arising from the use of any of the criteria can be identified and eliminated.

---

[8] Standard 13.7 states, "In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision." *Joint Standards, supra* note 3, at p. 146.

[9] *Joint Standards, supra* note 3, at p. 141.

[10] *Joint Standards, supra* note 3, at p. 141. Many test developers also caution against using their tests as the sole criterion in making a decision with high-stakes consequences for students. Discussion of this issue can be found in interpretive guides from test publishers, such as Riverside Publishing, Harcourt Brace, CTB McGraw Hill, and the Educational Testing Service, regarding the use of tests.

[11] National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, p. 3 (Jay P. Heubert & Robert M. Hauser eds., 1999) (hereinafter *High Stakes*).

[12] *See* regulations implementing Title VI of the Civil Rights Act of 1964, 34 C.F.R. §§ 100.3(a), 100.3(b)(1)(i) and (vi), 100.3(b)(2); regulations implementing Section 504 of the Rehabilitation Act of 1973, 34 C.F.R. §§ 104.4(a), 104.4(b)(1)(i) and (iv), 104.4(b)(4); regulations implementing Title IX of the Education Amendments of 1972, 34 C.F.R. §§ 106.31(a), 106.31(b).

Finally, this guide focuses primarily on tests used in making high-stakes decisions at the elementary and secondary education level. However, it is important to recognize that the general principles of sound educational measurement apply equally to tests used at the post-secondary education level, including admissions and other types of tests.[13] For example, post-secondary admissions policies and practices should be derived from and clearly linked to an institution's overarching educational goals, and the use of tests in the admissions process should serve those institutional goals.[14]

> **Standardized tests ... offer important benefits that should not be overlooked. ... Both the SAT [I] and ACT cover relatively broad domains that most observers would likely agree are relevant to the ability to do college work. Neither, however, measures the full range of abilities that are needed to succeed in college; important attributes not measured include, for example, persistence, intellectual curiosity, and writing ability. Moreover, these tests are neither complete nor precise measures of 'merit'—even academic merit.**
>
> National Research Council, *Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions*, pp. 21-22 (Alexandra Beatty, M.R.C. Greenwood & Robert L. Linn eds., 1999).

II. Foundations of the Resource Guide

A. Professional Standards of Sound Testing Practices

Chapter One summarizes the leading professionally recognized standards of sound testing practices within the educational measurement field. They include those described in the *Joint Standards*, which represents the primary statement of professional consensus regarding educational testing. Other leading professionally recognized standards of sound testing practices within the educational measurement field include the *Code of Fair Testing Practices in Education* (1988) and the *Code of Professional Responsibilities in Educational Measurement* (1995). The guide also cites recent reports from the NRC's Board on Testing

> **The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use and also can provide a route to broader and more equitable access to education. ... The improper use of tests, however, can cause considerable harm to test takers and other parties affected by test-based decisions.**
>
> American Educational Research Association, American Psychological Association & National Council on Measurement in Education, *Standards of Educational and Psychological Testing*, Introduction, p. 1 (1999).

---

[13] For additional information regarding testing at the post-secondary level, *see, e.g., Joint Standards, supra* note 3, at pp. 142-143; National Research Council, *Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions* (Alexandra Beatty, M.R.C. Greenwood & Robert L. Linn eds., 1999) (hereinafter *Myths and Tradeoffs*); *Educational Measurement* (Robert L. Linn ed., 3rd ed. 1989); *Ability Testing: Uses, Consequences, and Controversies*, Chapter 5 (Alexandra K. Wigdor & Wendell R. Garner eds., 1982).

[14] *Myths and Tradeoffs, supra* note 13, at p. 1.

BEST COPY AVAILABLE

and Assessment, including: *High Stakes: Testing for Tracking, Promotion and Graduation* (*High Stakes*, 1999); *Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions* (*Myths and Tradeoffs*, 1999); *Testing, Teaching, and Learning: A Guide for States and School Districts* (*Testing, Teaching, and Learning*, 1999); *Improving Schooling for Language-Minority Children: A Research Agenda* (*Improving Schooling for Language-Minority Children*, 1997); and *Educating One & All: Students with Disabilities and Standards-Based Reform* (*Educating One & All*, 1997).[15] These reports help explain or elaborate on principles that are stated in the *Joint Standards*.

Designed to provide criteria for the evaluation of tests, testing practices, and the effects of test use, the *Joint Standards* recommends that all professional test developers, sponsors, publishers, and users make efforts to observe the *Joint Standards* and encourage others to do so.[16] The *Joint Standards* includes chapters on the test development process (with a focus primarily on the responsibilities of test developers), the specific uses and applications of tests (with a focus primarily on the responsibilities of test users), and the rights and responsibilities of test takers. Because the *Joint Standards* is the most widely accepted collection of professional standards that is relied upon in developing testing instruments, this guide includes a discussion of specific standards that are contained within the *Joint Standards*, where relevant. Numbered standards that are referenced throughout this guide refer to specific standards contained within the *Joint Standards*.

To ensure that information presented in this guide is readable and accessible to educators and policy-makers, we have paraphrased language from relevant standards. Our goal in paraphrasing is to be concise and accurate. Where we have paraphrased in the text, we have also provided the full text of the relevant standards in the footnotes. Because the *Joint Standards* provides additional relevant discussion, we always encourage readers also to review the full document.

Professional test measurement standards provide important information that is relevant to making determinations about appropriate test use. The *Joint Standards* provides a frame of reference to assist in the evaluation of tests, testing practices, and the effects of test use. The *Joint Standards* cautions that the acceptability of a test or test application does not rest on the literal satisfaction of every standard in the *Joint Standards* and cannot be determined by using a checklist.[17] The exercise of professional judgment is a critical element in the interpretation and application of the standards, and the interpretation of individual standards should be considered in the overall context of the use of the test in question.[18] Finally, while the *Joint Standards* and federal nondiscrimination laws are closely aligned and mutually reinforcing, the failure to meet a particular professional test measurement standard does not necessarily constitute a lack of compliance with federal civil rights laws. Conversely, compliance with professional test measurement standards does not necessarily constitute compliance with all applicable federal civil rights laws.

---

[15] The National Resource Council of the National Academy of Sciences, which is an independent, private, nonprofit entity, established the NRC's Board on Testing and Assessment in 1993 to help policy-makers evaluate the use of tests, alternative assessments, and other indicators commonly used as tools of public policy. The Board provides guidance for judging the quality of testing or assessment technologies and the intended and unintended consequences of particular uses of these technologies. The Board concentrates on topics and conducts activities that serve the general public interest.

[16] *Joint Standards, supra* note 3, at Introduction, p. 2.

[17] *Joint Standards, supra* note 3, at Introduction, p. 4.

B. Legal Principles

Chapter Two of the guide discusses the federal constitutional, statutory, and regulatory nondiscrimination principles that apply to the use of tests for high-stakes purposes. This guide is intended to reflect existing legal principles and does not establish new federal legal requirements. The primary legal focus of the resource guide is an explanation of principles that are clearly embedded in four nondiscrimination laws that have been enacted by Congress: Title VI of the Civil Rights Act of 1964 (Title VI), Title IX of the Education Amendments of 1972 (Title IX), Section 504 of the Rehabilitation Act of 1973 (Section 504), and Title II of the Americans with Disabilities Act of 1990 (Title II).[19] Within the U.S. Department of Education, the Office for Civil Rights has responsibility for enforcing the requirements of these four statutes and their implementing regulations. The due process and equal protection requirements of the Fifth and Fourteenth Amendments to the U.S. Constitution have also been applied by courts to issues regarding the use of tests in making high-stakes educational decisions. Although the Office for Civil Rights does not enforce federal constitutional provisions, a brief overview of these fundamental constitutional principles has been included to provide educators with a more complete picture of relevant legal standards.

III. Basic Principles

The brief overview of the test measurement and legal principles that follows establishes the framework for more detailed discussions of test quality in Chapter One and federal legal standards in Chapter Two.

---

[18] *Joint Standards, supra* note 3, at Introduction, p. 4.

[19] Title VI prohibits discrimination on the basis of race, color, and national origin by recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Title VI is found at 34 C.F.R. Part 100. Title IX prohibits discrimination on the basis of sex by recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Title IX is found at 34 C.F.R. Part 106. Section 504 prohibits discrimination on the basis of disability by recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Section 504 is found at 34 C.F.R. Part 104. Title II prohibits discrimination on the basis of disability by public entities, regardless of whether they receive federal funding. The U.S. Department of Justice's regulation implementing Title II is found at 28 C.F.R. Part 35.

A. Test Use Principles

1. Educational Objectives and Context

> **Decisions about tracking, promotion, and graduation differ from one another in important ways. They differ most importantly in the role that mastery of past material and readiness for new material play.**
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, p. 4 (Jay P. Heubert & Robert M. Hauser eds., 1999).

Tests that are used in educationally appropriate ways and that are valid for the purposes used can serve as important instruments to help educators do their job. Before any state, school district, or educational institution administers a test, the objectives for using the test should be clear: What are the intended goals for and uses of the test in question? As an educational matter, the answer to this question will guide all other relevant inquiries about whether the test use is educationally appropriate. The context in which a test is to be administered, the population of test takers, the intended purpose for which the test will be used, and the consequences of such use are important considerations in determining whether the test would be appropriate for a specific type of decision, including placement, promotion, or graduation decisions.

Once education agencies or institutions have determined the underlying goals they want to accomplish, they need to identify the types of information that will best inform their decision-making. Information may include test results and other relevant measures that will be able to accurately and fairly address the purpose specified by the agencies or institutions.[20] When test results are used as part of high-stakes decision-making about student promotion or graduation, students should be given a reasonable number of opportunities to demonstrate mastery,[21] and students should have had an adequate opportunity to learn the material being tested.[22]

---

[20] *See* Standard 13.7 (n.8) in *Joint Standards, supra* note 3, at p. 146.

[21] Standard 13.6 states, "Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on equivalent forms of the test or be provided with construct-equivalent testing alternatives of equal difficulty to demonstrate the skills or knowledge. In most circumstances, when students are provided with multiple opportunities to demonstrate mastery, the time interval between the opportunities should allow for students to have the opportunity to obtain the relevant instructional experiences." *Joint Standards, supra* note 3, at p. 146.

[22] Standard 13.5 states, "When test results substantially contribute to making decisions about student promotion or graduation, there should be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn." *Joint Standards, supra* note 3, at p. 146.

BEST COPY AVAILABLE

a. Placement Decisions

> **[At the elementary and secondary education level,] appropriate test use for ... all students requires that their scores not lead to decisions or placements that are educationally detrimental.**
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, pp.40-41 (Jay P. Heubert & Robert M. Hauser eds., 1999).

Placement decisions are by their very nature used to make a decision about the future. Tests used in placement decisions generally determine what kinds of programs, services, or interventions will be most appropriate for particular students. Decisions concerning the appropriate educational program for a student with a disability, placement in gifted and talented programs, and access to language services are examples of placement decisions. The *Joint Standards* states that there should be adequate evidence documenting the relationship among test scores, appropriate instructional programs, and desired student outcomes.[23] When evidence about the relationship is limited, the test results should usually be considered in light of other relevant student information.[24]

b. Promotion Decisions

> **Neither a test score or any other kind of information can justify a bad decision. Research shows that students are typically hurt by simple retention and repetition of a grade in school without remedial and other instructional support services. In the absence of effective services for low-performing students, better tests will not lead to better educational outcomes.**
>
> National Research Council, *High-Stakes: Testing for Tracking, Promotion, and Graduation*, p. 3 (Jay P. Heubert & Robert M. Hauser eds., 1999).

Student promotion decisions are generally viewed as decisions incorporating a determination about whether a student has mastered the subject matter or content of instruction provided to the student and a determination regarding whether the student will be able to master the content at the next grade level (a placement decision).[25] When a test given for promotion purposes is being used to certify mastery, the use of the test should adhere to professional standards for certifying knowledge and

---

[23] Standard 13.9 states, "When test scores are intended to be used as part of the process for making decisions for educational placement, promotion, or implementation of prescribed educational plans, empirical evidence documenting the relationship among particular scores, the instructional programs, and desired student outcomes should be provided. When adequate empirical information is not available, users should be cautioned to weigh the test results accordingly in light of other relevant information about the student." *Joint Standards, supra* note 3, at p. 147.

[24] Standard 13.9 (n.23) in *Joint Standards, supra* note 3, at p. 147.

[25] *High Stakes, supra* note 11, at p. 123.

BEST COPY AVAILABLE

skills for all students.[26] As indicated in the *Joint Standards*, it is important that there "be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn."[27] Educational institutions should have information indicating an alignment among the curriculum, instruction, and material covered on such a test used for high-stakes purposes. To the extent that a test for promotion purposes is being used as a placement device, it should also adhere, as appropriate, to professional standards regarding tests used for placement purposes.[28]

c. Graduation Decisions

Graduation decisions are generally certification decisions: The diploma certifies that the student has reached an acceptable level of mastery of knowledge and skills.[29] When large-scale standardized tests are used in making graduation decisions, as indicated in the *Joint Standards*, there should "be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn."[30] Therefore, all students should be provided a meaningful opportunity to acquire the knowledge and skills that are being tested, and information should indicate an alignment among the curriculum, instruction, and material covered on the test used as a condition for graduation.[31]

2. Overarching Principles

In the elementary and secondary education context, regardless of whether tests are being used to make placement, promotion, or graduation decisions, the NRC's Board on Testing and Assessment has identified three principle criteria, based on established professional standards, that can help inform and guide conclusions regarding the appropriateness of a particular test use.[32]

---

[26] *See* Standard 13.5 (n.22) and 13.6 (n.21) in *Joint Standards, supra* note 3, at p. 146; *High Stakes, supra* note 11, at p. 123.

[27] Standard 13.5 (n.22) in *Joint Standards, supra* note 3, at p. 146; *see also High Stakes, supra* note 11 at pp. 124-125.

[28] *See* Standard 13.2 and 13.9 (n.23) in *Joint Standards, supra* note 3, at pp. 145, 147; *see also High Stakes, supra* note 11, at p. 123.

Standard 13.2 states, "In educational settings, when a test is designed or used to serve multiple purposes, evidence of the test's technical quality should be provided for each purpose." *Joint Standards, supra* note 3, at p. 145.

[29] *High Stakes, supra* note 11, at p. 166.

[30] Standard 13.5 (n.22) in *Joint Standards, supra* note 3, at p. 146.

[31] Sometimes scores from a test used for graduation purposes are used to provide remediation instruction for students who do not pass the test. In this case, "[s]chools that give graduation tests early . . . assume that such tests are diagnostic and that students who fail can benefit from effective remedial instruction . . . Using these test results to place a pupil in a remedial class or other intervention also involves a prediction about the student's performance--that is, that as a result of the placement, the student's mastery of the knowledge and skills measured by the test will improve. Thus, evidence that a particular treatment (in this case, the remedial program) benefits students who fail the test would be an appropriate part of the test validation process." *High Stakes, supra* note 11, at p. 171.

(1) *Measurement validity: Is a test valid for a particular purpose, and does it accurately measure the test taker's knowledge in the content area being tested?*

State and local education agencies and educational institutions should ensure that a test actually measures what it is intended to measure for all students. The inferences derived from the test scores for a given use — for a specific purpose, in a specific type of situation, and with specific types of students — are validated, rather than the test itself. It is important for educators who use the test to obtain adequate evidence of test quality (including validity and reliability evidence), evaluate the evidence, and ensure that the test is used appropriately in a manner that is consistent with information provided by the developers or through supplemental validation studies.

(2) *Attribution of cause: Does a student's performance on a test reflect knowledge and skills based on appropriate instruction, or is it attributable to poor instruction or to such factors as language barriers unrelated to the skills being tested?*

In some contexts, whether a particular test use is appropriate depends on whether test scores are an accurate reflection of a student's knowledge or skills or whether they are influenced by extraneous factors unrelated to the specific skills being tested. For example, when tests are used in making student promotion or graduation decisions, state and local education agencies should ensure that all students have an equal opportunity to acquire the knowledge and skills that are being tested.[33] In some situations, it may be necessary to provide appropriate accommodations for limited English proficient students and students with disabilities to accurately and effectively measure students' knowledge and skills in the particular content area being assessed.[34]

---

[32] *High Stakes, supra* note 11, at p. 23 (citing National Research Council, *Placing Children in Special Education: A Strategy for Equity* (1982)).

[33] Standard 7.10 states, "When the use of a test results in outcomes that affect the life chances or educational opportunities of examinees, evidence of mean test score differences between relevant subgroups of examinees should, where feasible, be examined for subgroups for which credible research reports mean differences for similar tests. Where mean differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct underrepresentation or construct-irrelevant variance. While initially, the responsibility of the test developer, the test user bears responsibility for uses with groups other than those specified by the developer." *Joint Standards, supra* note 3, at p. 83.

[34] *See Joint Standards, supra* note 3, at pp. 91-106.

(3) *Effectiveness of treatment: Do test scores lead to placements and other consequences that are educationally beneficial?*

The most basic obligation of educators at the elementary and secondary school levels is to meet the needs of students as they find them, with their different backgrounds, and to teach knowledge and skills to allow them to grow to maturity with meaningful expectations of a productive life in the workforce and elsewhere.[35] This obligation regarding elementary and secondary education is no less present when educators administer tests and evaluate and act on students' test results than it is during classroom instruction. Recognizing that tests used in the education setting should be integral to the learning and achievement of students, one federal court distinguished between testing in the employment and education settings:

> If tests predict that a person is going to be a poor employee, the employer can legitimately deny the person the job, but if tests suggest that a young child is probably going to be a poor student, a school cannot on that basis alone deny that child the opportunity to improve and develop the academic skills necessary to success in our society.[36]

Tests, in short, should be instruments used by elementary and secondary educators to help students achieve their full potential. Test scores should lead to consequences that are educationally beneficial for students. When making high-stakes decisions that involve the use of tests, it is important for policy-makers and educators to consider the intended and unintended consequences that may result from the use of the test scores.[37]

---

[35] *See Brown v. Board of Educ.*, 347 U.S. 483, 493 (1954) (stating that "[education] is required in the performance of our most basic public responsibilities, . . . is the very foundation of good citizenship, . . . [and] is [a] principal instrument . . . in preparing [the child] for later professional training . . . .").

[36] *Larry P. v. Riles*, 793 F.2d 969, 980 (9th Cir. 1984) (quoting *Larry P. v. Riles*, 495 F. Supp. 926, 969 (N.D. Cal. 1979)).

[37] For example, research indicates that students in low-track classes often do not have the opportunity to acquire knowledge and skills strongly associated with future success that is offered to students in other tracks. The National Research Council recommends that neither test scores nor other information should be used to place students in such classes. *High Stakes*, *supra* note 11, at p. 282.

> These criteria [measurement validity, attribution of cause, and effectiveness of treatment], based on established professional standards, lead to the following basic principles of appropriate test use for educational decisions:
>
> - The important thing about a test is not its validity in general, but its validity when used for a specific purpose. Thus, tests that are valid for influencing classroom practice, "leading" the curriculum, or holding schools accountable are not appropriate for making high-stakes decisions about individual student mastery unless the curriculum, the teaching, and the test(s) are aligned.
> - Tests are not perfect. Test questions are a sample of possible questions that could be asked in a given area. Moreover, a test score is not an exact measure of a student's knowledge or skills. A student's score can be expected to vary across different versions of a test – within a margin of error determined by the reliability of the test – as a function of the particular sample of questions asked and/or transitory factors, such as the student's health on the day of the test. Thus, no single test score can be considered a definitive measure of a student's knowledge.
> - An educational decision that will have a major impact on a test taker should not be made solely or automatically on the basis of a single test score. Other relevant information about the student's knowledge and skills should also be taken into account.
> - Neither a test score nor any other kind of information can justify a bad decision. Research shows that students are typically hurt by simple retention and repetition of a grade in school without remedial and other instructional supports. In the absence of effective services for low-performing students, better tests will not lead to better educational outcomes.
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion and Graduation*, p. 3 (Jay P. Heubert & Robert M. Hauser eds., 1999).

B. Legal Principles

Federal constitutional, statutory, and regulatory principles form the federal legal nondiscrimination framework applicable to the use of tests for high-stakes purposes. Title VI, Title IX, Section 504, and Title II, as well as the equal protection clause of the Fourteenth Amendment to the United States Constitution, prohibit intentional discrimination based on race, national origin, sex, or disability.[30] In addition, the regulations that implement Title VI, Title IX, Section 504, and Title II prohibit intentional discrimination as well as policies or practices that have a discriminatory disparate impact on students based on their race, national origin, sex, or disability.[39] The Section 504 regulation and the Individuals with Disabilities Education Act (IDEA) contain specific provisions relevant to the use of high-stakes tests for individuals with disabilities.[40]

---

[30] The United States Supreme Court has held that "Title VI itself directly reached only instances of intentional discrimination . . . [but that] actions having an unjustifiable disparate impact on minorities could be addressed through agency regulations designed to implement the purposes of Title VI." *Alexander v. Choate*, 439 U.S. 287, 295 (1985), discussing *Guardians Ass'n v. City Service Comm'n of N.Y.*, 403 U.S. 582 (1983). The United States Supreme Court has never expressly ruled on whether Section 504, Title II and Title IX statutes prohibit not only intentional discrimination, but, unlike Title VI, prohibit disparate impact discrimination as well. *See, e.g., Choate*, 409 U.S. at 294-97 & n.11 (observing that Congress might have intended the Section 504 statute itself to prohibit disparate impact discrimination). Section 504 and Title II require reasonable modifications where necessary to enable persons with disabilities to participate in or enjoy the benefits of public services. Regardless, the regulations implementing Section 504, Title II, and Title IX, like the Title VI regulation, explicitly prohibit actions having discriminatory effects as well as actions that are intentionally discriminatory.

These sources of legal authority should be considered in conjunction with the test measurement principles discussed in this guide to ensure that standardized tests are used in a manner that supports sound educational decisions, regardless of the race, national origin (including limited English proficiency), sex, or disability of the students affected. Some of the issues that have been considered by federal courts in assessing the legality of specific testing practices for making high-stakes decisions include:[41]

- The use of an educational test for a purpose for which the test was not designed or validated;
- The use of a test score as the sole criterion for the educational decision;
- The nature and quality of the opportunity provided to students to master required content, including whether classroom instruction included the material covered by a test administered to determine student achievement;
- The significance of any fairness problems identified, including evidence of differential prediction criterion and possible cultural biases in the test or in test items; and
- The educational basis for establishing passing or cutoff scores.

1. Frameworks for Analysis

a. Different Treatment

Under federal law, policies and practices generally must be applied consistently to similarly situated individuals or groups regardless of their race, national origin, sex, or disability.

---

[39] 34 C.F.R. § 100.3(b)(2) (Title VI); 34 C.F.R. §§ 106.21(b)(2), 106.36(b), 106.52 (Title IX); 34 C.F.R. § 104.4(b)(4)(i) (Section 504); 28 C.F.R. § 35.130(b)(3) (Title II).

The authority of federal agencies to issue regulations with an "effects" standard has been consistently acknowledged by U.S. Supreme Court decisions and applied by lower federal courts addressing claims of discrimination in education. *See, e.g., Choate*, 469 U.S. at 289-300; *Guardians Ass'n*, 463 U.S. at 584-93; *Lau v. Nichols*, 414 U.S. 563, 568 (1974); *see also* Memorandum from the Attorney General for Heads of Departments and Agencies that Provide Federal Financial Assistance, *Use of the Disparate Impact Standard in Administrative Regulations under Title VI of the Civil Rights Act of 1964* (July 14, 1994).

[40] The Individuals with Disabilities Education Act (IDEA) establishes rights and protections for students with disabilities and their families. It also provides federal funds to local school districts and state agencies to assist in educating students with disabilities. *See* 20 U.S.C. §§ 1400(1)(c) *et seq*. The specific sections of the regulations implementing Section 504 and the IDEA bearing on testing are 20 U.S.C. §§ 1412(a)(17), 1414(b); 34 C.F.R. §§ 104.4(b)(4), 104.33, 104.35, 104.42(b), 104.44, 300.138 - .139, 300.530 - .536.

[41] For specific court decisions examining these issues, see discussion *infra* Chapter 2 (Legal Principles) & nn.167-171.

For example, a court concluded that a school district had intentionally treated students differently on the basis of race where minority students whose test scores qualified them for two or more ability levels were more likely to be assigned to the lower level class than similarly situated white students, and no explanatory reason was evident.[42]

In addition, educational systems that previously discriminated by race in violation of the Fourteenth Amendment and have not achieved unitary status have an obligation to dismantle their prior *de jure* segregation. In such instances, school districts are under "a 'heavy burden' of showing that actions that [have] increased or continued the effects of the dual system serve important and legitimate ends."[43] When such a school district or educational system uses a test or assessment procedure for a high-stakes purpose that has significant racially disparate effects, to justify the test use, the school district must show that the test results are not due to the present effects of prior segregation or that the practice or procedure remedies the present effects of such segregation by offering better educational opportunities.[44]

b. Disparate Impact

The federal nondiscrimination regulations also provide that a recipient of federal funds may not "utilize criteria or methods of administration which have the effect of subjecting individuals to discrimination."[45] Thus, discrimination under federal law may occur where the application of neutral criteria has disparate effects and those criteria are not educationally justified.

---

[42] *See People Who Care v. Rockford Bd. of Educ.*, 851 F. Supp. 905, 958-1001 (N.D. Ill. 1994), *remedial order rev'd, in part*, 111 F.3d 528 (7th Cir. 1997). On appeal, the Seventh Circuit Court of Appeals stated that the appropriate remedy based on the facts in this case was to require the district to use objective, non-racial criteria to assign students to classes, rather than abolishing the district's tracking system. *See id.* at 536.

[43] *Dayton Bd. of Educ. v. Brinkman*, 443 U.S. 526, 538 (1979) (quoting *Green v. County School Bd.*, 391 U.S. 430, 439 (1968)).

[44] *See Debra P. v. Turlington*, 644 F.2d 397, 407 (5th Cir. 1981) ("[Defendants] failed to demonstrate either that the disproportionate failure [rate] of blacks was not due to the present effects of past intentional segregation or, that as presently used, the diploma section was necessary [in order] to remedy those effects."); *McNeal v. Tate County Sch. Dist.*, 508 F.2d 1017, 1020 (5th Cir. 1975) (ability grouping method that causes segregation may nonetheless be used "if the school district can demonstrate that its assignment method is not based on the present results of past segregation or that the method of assignment will remedy such effects through better educational opportunities"); *see also United States v. Fordice*, 505 U.S. 717, 731 (1992) ("If the State [university system] perpetuates policies and practices traceable to its prior system that continue to have segregative effects . . . and such policies are without sound educational justification and can be practically eliminated, the State has not satisfied its burden of proving that it has dismantled its prior system."); *cf. GI Forum v. Texas Educ. Agency*, 87 F. Supp. 2d 667, 673, 684 (W.D. Tex. 2000) (the court concluded, based on the facts presented, that the test seeks to identify inequities and address them; the state had ensured that the exam is strongly correlated to material actually taught in the classroom; remedial efforts, on balance, are largely successful; and minority students have continued to narrow the passing gap at a rapid rate).

[45] 34 C.F.R. § 100.3(b)(2) (Title VI); 34 C.F.R. § 104.4(b)(4)(i) (Section 504); 28 C.F.R. § 35.130(b)(3)(i) (Title II); *see also* 34 C.F.R. §§ 106.21, 106.31, 106.36(b), 106.52 (Title IX). In *Guardians Association*, the United States Supreme Court upheld the use of the effects test, stating that the Title VI regulation forbids the use of federal funds "not only in programs that intentionally discriminate, but also in those endeavors that have a [racially disproportionate] impact on racial minorities." 463 U.S. at 589.

It is ... important to note that group differences in test performance do not necessarily indicate problems in a test, because test scores may reflect real differences in achievement. These, in turn, may be due to a lack of access to a high quality curriculum and instruction. Thus, a finding of group differences calls for a careful effort to determine their cause.

National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, p. 5 (Jay P. Heubert & Robert M. Hauser eds., 1999).

The disparate impact analysis has been frequently misunderstood to indicate a violation of law based merely on disparities in student performance and to obligate educational institutions to change their policies and procedures to guarantee equal results. Under federal law, a statistically significant difference in outcomes creates the need for further examination of the educational practices that have caused the disparities in order to ensure accurate and nondiscriminatory decision-making, but disparate impact alone is not sufficient to prove a violation of federal civil rights laws.

Courts applying the disparate impact test have generally examined three questions to determine if the practice at issue is discriminatory: (1) Does the practice or procedure in question result in significant differences in the award of benefits or services based on race, national origin, or sex? (2) Is the practice or procedure educationally justified? (3) Is there an equally effective alternative that can accomplish the institution's educational goal with less disparity?[46] (For a discussion of disability discrimination, including disparate impact discrimination, see discussion *infra* Chapter 2 (Legal Principles) Part III (Testing Students with Disabilities).[47])

Under the disparate impact analysis, the party challenging the test has the burden of establishing disparate impact, generally through evidence of a statistically significant difference in the awards of benefits or services. If disparate impact is established, the educational institution must demonstrate the educational justification (also referred to as "educational necessity") for the practice in question.[48] If sufficient evidence of an

---

[46] Courts use a variety of terms when discussing whether an alternative offered by the party challenging the practice would effectively further the institution's goals. *See, e.g.*, *Georgia State Conf. of Branches of NAACP v. Georgia*, 775 F.2d 1403, 1417 (11th Cir. 1985) (party challenging the practice "may ultimately prevail by proffering an equally effective alternative practice which results in less racial disproportionality"); *Elston v. Talladega*, 997 F.2d 1394, 1407 (11th Cir. 1993) (party challenging the practice "will still prevail if able to show that there exists a comparably effective alternative practice which would result in less disproportionality"). These terms ("equally effective" and "comparably effective") appear to be used synonymously.

[47] Disparate impact disability discrimination may take forms that are not always amenable to analysis through the three-part approach used in race and sex discrimination cases. For example, statistical proof may not be necessary when evaluating the effects of architectural barriers. *See Choate*, 469 U.S. at 297-300. For this reason, disability discrimination is discussed separately in this guide. *See* discussion *infra* Chapter 2 (Legal Principles) Part III (Testing of Students with Disabilities).

[48] *Elston*, 997 F.2d at 1412.

BEST COPY AVAILABLE

educational justification has been provided, the party challenging the test must then establish, in order to prevail, that an alternative practice with less disparate impact is equally effective in furthering the institution's educational goals.[49]

2. Principles Relating to Inclusion and Accommodations

a. Limited English Proficient Students

The obligations of states and school districts with regard to testing of limited English proficient students for high-stakes purposes in elementary and secondary schools must be examined within the overall context of the Title VI obligation to provide equal educational opportunities to limited English proficient students. Under Title VI, school districts have an obligation to identify limited English proficient students and to provide them with an instructional program or services that enables them to acquire English-language proficiency as well as the knowledge and skills that all students are expected to master.[50] School districts also have a responsibility to ensure that the instructional program or services provide limited English proficient students with a meaningful opportunity to acquire the academic knowledge and skills covered by tests required for graduation or other educational benefits.

In addition, states or school districts using tests for high-stakes purposes must ensure that, as with all students, the tests effectively measure limited English proficient students' knowledge and skills in the particular content area being assessed. For limited English proficient elementary and secondary school students in particular, it may be necessary in some situations to provide accommodations so that the tests provide accurate information about the knowledge and skills intended to be measured.[51]

b. Students with Disabilities

Under Section 504, Title II, and the IDEA,[52] school districts have a responsibility to provide elementary and secondary school students with disabilities with a free appropriate public education. Providing effective instruction in the general curriculum for students with disabilities is an important aspect of providing a free appropriate public education. Under federal law, students with disabilities must be included in statewide or districtwide assessment programs and provided with appropriate accommodations, if necessary.[53] There must be an individualized determination of whether a student with a disability will participate in a particular test and the appropriate accommodations, if any, that a student with a disability will need. This individualized determination must be addressed through the individualized education program (IEP) process or other applicable evaluation procedures and included in either the student's IEP or Section 504 plan.[54] The IDEA also requires state or local education agencies to develop guidelines for the relatively small number of students with disabilities who cannot take part in statewide or districtwide tests to participate in alternate assessments.[55]

---

[49] *Georgia State Conf.*, 775 F.2d at 1417; *see also* Department of Justice, *Title VI Legal Manual*, p. 2.

[50] *See* Equal Educational Opportunities Act of 1974, 20 U.S.C. §§ 1701-1720; *Lau*, 414 U.S. at 568-69; *Castaneda v. Pickard*, 648 F.2d 989, 1011 (5th Cir. 1981); Michael L. Williams, Former Assistant Secretary for Civil Rights, *Memorandum to OCR Senior Staff* (September 27, 1991) (hereinafter *Williams Memorandum*).

[51] States and school districts are also required to provide limited English proficient students with "reasonable adaptations and accommodations" in certain situations when using assessments for the purpose of holding schools and districts accountable for student performance under Title I. Title I of the Elementary and Secondary Education Act, 20 U.S.C. § 6311(b)(3)(F)(ii). Moreover, Title I requires States, to the extent practicable, to provide native-language assessments to LEP students for Title I accountability purposes if that is the language and form of assessment most likely to yield accurate and reliable information about what students know and can do. 20 U.S.C. § 6311(b)(3)(F)(iii). For a discussion of comparability issues arising in the testing of LEP students, *see* discussion *infra* Chapter 2 (Legal Principles) Part II (Testing of Students with Limited English Proficiency).

[52] The Section 504 regulation is found at 34 C.F.R. Part 104. The Title II regulation is found at 28 C.F.R. Part 35. The IDEA regulation is found at 34 C.F.R. Part 300.

Finally, under Section 504, post-secondary education institutions may not make use of any test or criterion for admission that has a disproportionate adverse impact on individuals with disabilities unless (1) the test or criterion, as used by the institution, has been validated as a predictor of success in the education program or activity and (2) alternate tests or criteria that have a less disproportionate adverse impact are not shown to be available by the party asserting that the test or criterion is discriminatory.[56] Admissions tests must be selected and administered so as best to ensure that, when a test is administered to an applicant with a disability, the test results accurately reflect the applicant's aptitude or achievement level, rather than reflecting the effect of the disability (except where the functions impaired by the disability are the factors the test purports to measure).[57] A student requesting an accommodation must initially provide documentation of the disability and the need for accommodation. Admissions tests designed for persons with impaired sensory, manual, or speaking skills must be offered as often and in as timely a manner as are other admissions tests. Admissions tests also must be offered in facilities that, on the whole, are accessible to individuals with disabilities.

---

[53] States and school districts are also required to provide students with disabilities with "reasonable adaptations and accommodations" in certain situations when using assessments for the purpose of holding schools and districts accountable for student performance under Title I. 20 U.S.C. § 6311(b)(3)(F)(ii).

[54] Under the IDEA, students with disabilities must be included in state and districtwide assessment programs. 34 C.F.R. § 300.138(a). However, if the IEP team determines that a student should not participate in a particular statewide or districtwide assessment of student achievement (or part of such an assessment), the student's IEP must include statements of why that test is not appropriate for the student and how the student will be assessed. 34 C.F.R. § 300.347(a)(5). The IDEA also requires state or local education agencies to develop guidelines for students with disabilities who cannot take part in state- and districtwide assessments to participate in alternate assessments; these alternate assessments must be developed and conducted beginning not later than July 1, 2000. 34 C.F.R. § 300.138(b).

[55] 34 C.F.R. § 300.138(b).

[56] 34 C.F.R. § 104.42(b)(2).

[57] 34 C.F.R. § 104.42(b)(3).

3. Federal Constitutional Questions Related to the Use of Tests as Part of High-Stakes Decision-Making for Students

The equal protection and due process requirements of the Fifth and Fourteenth Amendments to the U.S. Constitution also apply to ensure that high-stakes decisions by public schools or states involving the use of tests are made appropriately.[58] The equal protection principles involved in discrimination cases are, generally speaking, the same as the standards applied to intentional discrimination (or different treatment) claims under the applicable federal nondiscrimination statutes.[59] Courts addressing due process claims have examined three questions related to the use of tests as bases for promotion or graduation decisions:

- Is the testing program reasonably related to a legitimate educational purpose?
- Have students received adequate notice of the test and its consequences?
- Have students actually been taught the knowledge and skills measured by the test?

Federal courts have typically deferred to educators' authority to formulate appropriate educational goals.[60] For example, improving the quality of education, ensuring that students can compete on a national and international level, and encouraging educational achievement through the establishment of academic standards have been found to be legitimate goals for testing programs.[61] The constitutional inquiry then proceeds to examine whether the challenged testing program is reasonably related to the educators' legitimate goals or whether the program is arbitrary and capricious or fundamentally unfair.[62]

---

[58] The requirements of Title VI, Title IX and Section 504 apply only to recipients of federal financial assistance. The protections afforded by the Fifth and Fourteenth Amendments to the U.S. Constitution apply to actions by "state actors" and are not dependent upon receipt of federal financial assistance.

[59] Federal cases may also involve equal protection challenges to a jurisdiction's use of tests in which the claim is not based on race or sex discrimination, but, instead, on assertions that the classifications made by the jurisdiction on the basis of test scores are unreasonable, regardless of the race or sex of the students affected. *See GI Forum*, 87 F. Supp. 2d at 682. As a general matter, courts express reluctance to second guess a state's educational policy choices when faced with such challenges, although they recognize that a state cannot "exercise that [plenary] power without reason and without regard to the United States Constitution." *Debra P.*, 644 F.2d at 403. When there is no claim of discrimination based on membership in a suspect class, the equal protection claim is reviewed under the rational basis standard. In these cases, the jurisdiction need show only that the use of the tests has a rational relationship to a valid state interest. *Id.* at 406; *Erik V. v. Causby*, 977 F. Supp. 384, 389 (E.D.N.C. 1997).

[60] *See Regents of the Univ. of Mich. v. Ewing*, 474 U.S. 214, 226-27 (1985); *Debra P.*, 644 F.2d at 406; *Anderson v. Banks*, 520 F. Supp. 472, 506 (S.D. Ga. 1981).

[61] *See Ewing*, 474 U.S. at 226-27; *Debra P.*, 644 F.2d at 406; *Anderson*, 520 F. Supp. at 506.

[62] *See Ewing*, 474 U.S. at 222, 226-27; *Debra P.*, 644 F.2d at 406; *GI Forum*, 87 F. Supp. 2d at 682; *Anderson*, 520 F. Supp. at 506.

In due process cases, courts have generally required advance notice of test requirements in order to give students a reasonable chance to understand the standards against which they will be evaluated and to learn the material for which they are to be accountable.[63] A reasonable transition period is required between the development of a new academic requirement and the attachment of high-stakes consequences to tests used to measure academic achievement. That time period varies, however, depending upon the precise context in which the high-stakes decision is to be made. Relevant inquiries affecting determinations about the constitutionality of notice and timing have included questions about the alignment of curriculum and instruction with material tested, the number of test taking opportunities provided to students, tutorial or remedial opportunities provided to students, and whether factors in addition to test scores can affect high-stakes decisions.

Finally, in due process cases, federal courts have required, as a matter of "fundamental fairness," that students have a reasonable opportunity to learn the material covered by the test where passing the test is a condition of receipt of a high school diploma or a condition for grade-to-grade promotion.[64] For the test to meaningfully measure student achievement, the test, the curriculum, and classroom instruction should be aligned.[65]

---

[63] *See Brookhart v. Illinois Bd. Of Educ.*, 697 F.2d 179, 185 (7th Cir. 1983); *Debra P.*, 644 F.2d at 404; *Erik V.*, 977 F. Supp. at 389-90; *Anderson*, 520 F. Supp. at 1410-12.

[64] *See Brookhart*, 697 F.2d at 184-87; *Debra P.*, 644 F.2d at 406; *GI Forum*, 87 F. Supp. 2d at 682; *Anderson*, 520 F. Supp. at 509.

[65] *Brookhart*, 697 F.2d at 184-87; *Debra P.*, 644 F.2d at 406; *Anderson*, 520 F. Supp. at 509. Insofar as due process cases may involve additional questions regarding the validity, reliability, and fairness of the test used to address the educational institution's stated purposes, these issues are discussed in the portions of the guide addressing discrimination under federal civil rights laws.

CHAPTER 1: Test Measurement Principles

This chapter explains basic test measurement standards and related educational principles for determining whether tests used as part of making high-stakes decisions for students provide accurate and fair information. As explained in Chapter Two below, federal court decisions have been informed and guided by professional test measurement standards and principles. Understanding professional test measurement standards can assist in efforts to use tests wisely and to comply with federal nondiscrimination laws.[66] This chapter is intended as a helpful discussion of how to understand test measurement concepts and their use. These are not specific legal requirements, but rather are foundations for understanding appropriate test use.

Educational institutions use tests to accomplish specific purposes based on their educational goals, including making placement, promotion, graduation, admissions, and other decisions. It is only after educational institutions have determined the underlying goal they want to accomplish that they can identify the types of information that will best inform their decision-making. That information may include test results as well as other relevant measures that can effectively, accurately, and fairly address the purposes and goals specified by the institutions.[67] As stated in the *Joint Standards*, "When interpreting and using scores about individuals or groups of students, consideration of relevant collateral information can enhance the validity of the interpretation, by providing corroborating evidence or evidence that helps explain student performance. . . . As the stakes of testing increase for individual students, the importance of considering additional evidence to document the validity of score interpretations and the fairness in testing increases accordingly."[68]

Although this guide focuses on the use of tests, policy-makers and educators need to consider the soundness and relevance of the entire high stakes decision-making process, including other information used in conjunction with test results.[69]

In using tests as part of high-stakes decision-making, educational institutions should ensure that the test will provide accurate results that are valid, reliable, and fair for all test takers. This includes obtaining adequate evidence of test quality about the current test being proposed and its use, evaluating the evidence, and ensuring that appropriate test use is based on adequate evidence.[70] When test results are used to make high-stakes decisions about student promotion or graduation, educational institutions should provide students with a reasonable number of opportunities to demonstrate mastery and ensure that there is evidence available that students have had an adequate opportunity to learn the material being tested.[71]

---

[66] *See, e.g., High Stakes, supra* note 11, at pp. 59-60.

[67] Among other considerations, institutions will determine if they want test score interpretations that are norm-referenced or criterion-referenced, or both. Norm-referenced means that the performances of students are compared to the performances of other students in a specified reference population; criterion-referenced indicates the extent to which students have mastered specific knowledge and skills.

[68] *Joint Standards, supra* note 3, at p. 141; *see also* Standard 13.7 (n.8) in *Joint Standards, supra* note 3, at p. 146.

[69] *Joint Standards, supra* note 3, at p. 141.

I. Key Considerations in Test Use

This section addresses the fundamental concepts of test validity and reliability. It will also discuss issues associated with ensuring fairness in the meaning of test scores, and issues related to using appropriate cut scores. Test developers and users as appropriate determine adequate validity and reliability, ensure fairness, and determine where to set and how to use cut scores appropriately for all students by accumulating evidence of test quality from relevant groups of test takers.

A. Validity

Test validity refers to a determination of how well a test actually measures what it says it measures. The *Joint Standards* defines validity as "[t]he degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test."[72] The demonstration of validity is multifaceted and must always be determined within the context of the specific use of a test. In order to promote readability, the discussion on validity presented here is meant to reflect this complex topic in an accurate, but concise and user-friendly way. The *Joint Standards* identifies and discusses in detail principles related to determining the validity of test results within the context of their use, and readers are encouraged to review the *Joint Standards*, Chapter 1, Validity, for additional, relevant discussion.[73]

---

[70] In order to provide educational institutions with tests that are accurate and fair, test developers should develop tests in accordance with professionally recognized standards, and provide educational institutions with adequate evidence of test quality.

Standard 1.4 states, "If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary." *Joint Standards, supra* note 3, at p. 18.

Standard 11.2 states, "When a test is to be used for a purpose for which little or no documentation is available, the user is responsible for obtaining evidence of the test's validity and reliability for this purpose." *Joint Standards, supra* note 3, at p. 113.

[71] *See* Standard 7.5, 13.5 (n.22) and 13.6 (n.21) in *Joint Standards, supra* note 3, at pp. 82, 146.

Standard 7.5 states, "In testing applications involving individualized interpretations of test scores other than selection, a test taker's score should not be accepted as a reflection of standing on the characteristic being assessed without consideration of alternate explanations for the test taker's performance on that test at that time." *Joint Standards, supra* note 3, at p. 82.

[72] *Joint Standards, supra* note 3, at pp. 9, 184.

[73] *Joint Standards, supra* note 3, at pp. 9-24.

There are three central points to keep in mind regarding validity:

- The focus of validity is not really on the test itself, but on the validity of the inferences drawn from the test results for a given use.
- All validity is really a form of "construct validity."
- In validating the inferences of the test results, it is important to consider the consequences of the test's interpretation and use.

1. Validity of the Inferences Drawn from the Scores

It is not the test that is validated per se, but the inferences or meaning derived from the test scores for a given use—that is, for a specific type of purpose, in a specific type of situation, and with specific groups of students. The meaning of test scores will differ based on such factors as how the test is designed, the types of questions that are asked, and the documentation that supports how all groups of students are interpreting what the test is asking and how effectively their performance can be generalized beyond the test.

For instance, in one case, the educational institution may want to evaluate how well students can analyze complex issues and evaluate implications in history. For a given amount of test time, they would want to use a test that measures the ability of students to think deeply about a few selected history topics. The meaning of the scores should reflect this purpose and the limits of the range of topics being measured on the test. In another case, the institution may want to assess how well students know a range of facts about a wide variety of historical events. The institution would want to use a test that measures a broad range of knowledge about many different occurrences in history. The inferences drawn from the scores should be validated to determine how well they measure students' knowledge of a broad range of historical facts, but not necessarily how well students analyze complex issues in history.

2. Construct Validity

Construct validity refers to the degree to which the scores of test takers accurately reflect the constructs a test is attempting to measure. The *Joint Standards* defines a construct as "the concept or the characteristic that a test is designed to measure."[74] Test scores and their inferences are validated to measure one or more constructs, which together comprise a particular content domain.[75] In K-12 education, these domains are often codified in state or district content standards covering various subject areas. For instance, the domain of mathematics as described in the state's elementary mathematics content standards may involve the constructs of mathematical problem-solving and knowledge of number systems. Items may be selected for a test that sample from this domain, and should be properly representative of the constructs identified within it. In that way, the meaning of the test scores should accurately reflect the knowledge and skills defined in the mathematics content standards domain.

---

[74] *Joint Standards*, *supra* note 3, at p. 173.

[75] The *Joint Standards* defines a content domain as "the set of behaviors, knowledge, skills, abilities, attitudes or other characteristics to be measured by a test, represented in a detailed specification, and often organized into categories by which items are classified." *Joint Standards*, *supra* note 3, at p. 174. A domain, then, represents a definition of a content area for the purposes of a particular test. Other tests will likely have a different definition of what knowledge and skills a particular content area entails.

Validity should be viewed as the overarching, integrative evaluation of the degree to which all accumulated evidence supports the intended interpretation of the test scores for a proposed purpose.[76] This unitary and comprehensive concept of validity is referred to as "construct validity." Different sources of validity evidence may illuminate different aspects of validity, but they do not represent distinct types of validity.[77]

Therefore, "construct validity" is not just one of the many types of validity—it *is* validity. The process of test validation "logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation for the proposed use."[78] Demonstrating construct validity then means gathering a variety of types of evidence to support the intended interpretations and uses of test scores. "The decision about what types of evidence are important for validation in each instance can be clarified by developing a set of propositions that support the proposed interpretation for the particular purpose of testing."[79] These propositions provide details that support the claims that, for a proposed use, the test validly measures particular skills and knowledge of the students being tested. For instance, if a test is designed to measure students' learning of material described in a district's science content standards, evidence that the test is properly aligned with these standards for the types of students taking the test would be a crucial component of the test's validity. When such evidence is in place, users of the test can correctly interpret high scores as indicators that students have learned the designated material and low scores as evidence that they have not.

All validity evidence and the interpretation of the evidence are focused on the basic question: Is the test measuring the concept, skill, or trait in question? Is it, for example, really measuring mathematical reasoning or reading comprehension for the types of students that are being tested? A variety of types of evidence can be used to answer this question—none of which provides a simple yes or no answer. The exact nature of the types of evidence that need to be accumulated is directly related to the intended use of the test, which includes evidence regarding the skills and knowledge being measured, evidence documenting validity for the stated purpose, and evidence of validity for all groups of students taking the test.[80]

---

[76] *See Joint Standards, supra* note 3, at pp. 9-11, 184.

[77] Therefore, construct validity can be seen as an umbrella that encompasses what has previously been described as predictive validity, content validity, criterion validity, discriminant validity, etc. Rather, these terms refer to types or sources of evidence that can be accumulated to support the validity argument. Definitions of these terms can be found in Appendix B, Measurement Glossary.

[78] *Joint Standards, supra* note 3, at p. 9.

[79] *Joint Standards, supra* note 3, at p. 9.

For instance, an educational institution may want to use a test to help make promotion decisions. It may also want to use a test to place students in the appropriate sequence of courses. In each situation, the types of validity evidence an institution would expect to see would depend on how the test is being used.

In making promotion decisions, the test should reflect content the student has learned. Appropriate validation would include adequate evidence that the test is measuring the constructs identified in the curriculum, and that the inferences of the scores accurately reflect the intended constructs for all test takers. Validation of the decision process involving the use of the test would include adequate evidence that low scores reflect lack of knowledge of students after they have been taught the material, rather than lack of exposure to the curriculum in the first place.

In making placement decisions, on the other hand, the test may not need to measure content that the student has already learned. Rather, at least in part, the educational institution may want the test to measure aptitude for the future learning of knowledge or skills that have been identified as necessary to complete a course sequence. Appropriate validation would include documentation of the relationship between what constructs are being measured in the test and what knowledge and skills are actually needed in the future placements. Evidence should also provide documentation that scores are not significantly confounded by other factors irrelevant to the knowledge and skills the test is intending to measure.

Institutions often think about using the same test for two or more purposes. This is appropriate as long as the validity evidence properly supports the use of the test for each purpose, and properly supports that the inferences of the results accurately reflect what the test is measuring for all students taking the test.[81]

The empirical evidence related to the various aspects of construct validity is collected throughout test development, during test construction, and after the test is completed. It is important for educators and policy-makers to understand and expect that the accumulated evidence spans the range of test development and implementation. There is not just one set of documentation collected at one point in time.[82]

---

[80] Rather than follow the traditional nomenclature (e.g. predictive validity, content validity, criterion validity, discriminant validity, etc.), the *Joint Standards* defines sources of validity evidence as evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on consequences of testing. *See Joint Standards, supra* note 3, at pp. 11-17.

[81] *See Joint Standards, supra* note 3, at pp. 9-24 (Chapter 1, Validity).

When the empirical database is large and includes results from a number of studies related to a given purpose, situation, and type of test takers, it may be appropriate to generalize validity findings beyond validity data gathered for one particular test use. That is, it may be appropriate to use evidence collected in one setting when determining the validity of the meaning of the test scores for a similar use. If the accumulated validity evidence for a particular purpose, situation, or subgroup is small, or features of the proposed use of the test differ markedly from an adequate amount of validity evidence already collected, evidence from this particular type of test use will generally need to be compiled.[83] Regardless of where the evidence is collected, educational institutions should expect adequate documentation of construct validity based on needs defined by the particular purposes and populations for which a test is being used.

When considering the types of construct validity evidence to collect, the *Joint Standards* emphasizes that it is important to guard against the two major sources of validity error. This error can distort the intended meaning of scores for particular groups of students, situations, or purposes.[84]

One potential source of error omits some important aspects of the intended construct being tested. This is called construct underrepresentation.[85] An example would be a test that is being used to measure English language proficiency. When the institution has defined English language proficiency as including specific skills in listening, speaking, reading, and writing the English language, and wants to use a test which measures these aspects, construct underrepresentation would occur if the test only measured the reading skills.

---

[82] Standard 3.6 states "The type of items, the response formats, scoring procedures; and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented." *Joint Standards, supra* note 3, at p. 44.

[83] As indicated in the *Joint Standards*, "The extent to which predictive or concurrent evidence of validity generalization can be used in new situations is in large measure a function of accumulated research. Although evidence of generalization can often help to support a claim of validity in a new situation, the extent of available data limits the extent to which the claim can be sustained." *Joint Standards, supra* note 3, at pp. 15-16.

[84] *Joint Standards, supra* note 3, at p. 10.

[85] Samuel Messick, *Validity*, in *Educational Measurement*, pp. 13-103 (Robert L. Linn ed., 3rd ed. 1989) (hereinafter Messick, *Validity*); Samuel Messick, *Validity of Psychological Assessment: Validations of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning*, American Psychologist 50(9), pp. 741-749 (September 1995) (hereinafter Messick, *Validity of Psychological Assessment*).

The other potential source of error occurs when a test measures material that is extraneous to the intended construct, confounding the ability of the test to measure the construct that it intends to measure. This source of error is called construct irrelevance.[86] For instance, how well a student reads a mathematics test may influence the student's subtest score in mathematics computation. In this case, the student's reading skills may be irrelevant when the skill of mathematics computation is what is being measured by the subtest.[87] Thus, in order to address considerations of construct underrepresentation and construct irrelevance it is important to collect evidence not only about what a test measures in particular types of situations or for particular groups of students, but also evidence that seeks to document that the intended meaning of the test scores is not unduly influenced by either of the two sources of validity error.

3. Considering the Consequences of Test Use

Evidence about the intended and unintended consequences of test use can provide important information about the validity of the inferences to be drawn from the test results, or it can raise concerns about an inappropriate use of a test where the inferences may be valid for other uses.

For instance, significant differences in placement test scores based on race, gender, or national origin may trigger a further inquiry about the test and how it is being used to make placement decisions.[88] The validity of the test scores would be called into question if the test scores are substantially affected by irrelevant factors that are not related to the academic knowledge and skills that the test is supposed to measure.[89]

On the other hand, a test may accurately measure differences in the level of students' academic achievement. That is, low scores may accurately reflect that some students do not know the content. However, test users should ensure that they interpret those scores

---

[86] Messick, *Validity, supra* note 85; Messick, *Validity of Psychological Assessment, supra* note 85.

[87] On the other hand, if an item is measuring the student's ability to apply mathematical skills in a written format (for instance when an item requires students to fill out an order form), then writing skills may not be extraneous to the construct being measured in this item.

[88] *See* Joint Committee on Testing Practices, *Code of Fair Testing Practices in Education* (1988).

[89] *See* Standard 1.24, 7.5 (n.71) and 7.6 in *Joint Standards, supra* note 3, at pp. 23-24, 82.

Standard 1.24 states, "When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or to the test's failure fully to represent the intended construct." *Joint Standards, supra* note 63, at p. 23.

Standard 7.6 states, "When empirical studies of differential prediction of a criterion for members of different subgroups are conducted, they should include regression equations (or an appropriate equivalent) computed separately for each group or treatment under consideration or an analysis in which the group or treatment variables are entered as moderator variables." *Joint Standards, supra* note 3, at p. 82.

> Standard 13.1
>
> **When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.**

correctly in the context of their high-stakes decisions.[90] For instance, test users could incorrectly conclude that the scores reflect lack of ability to master the content for some students when, in fact, the low test scores reflect the limited educational opportunities that the students have received. In this case, it would be inappropriate to use the test scores to place low-performing students in a special services program for students who have trouble learning and processing academic content.[91] It would be appropriate to use the test to evaluate program effectiveness, however.[92]

B. Reliability

Reliability refers to the degree of consistency of test results over test administrations, forms, items, scorers, and/or other facets of testing.[93] All indices of reliability are estimates of consistency, and all the estimates contain some error, since no test or other source of information is ever an "error-free" measure of student performance.[94] An example of reliability of test results over test administrations is when the same students, taking the test multiple times, receive similar scores. Consistency over parallel forms of a test occurs when forms are developed to be equivalent in content and technical characteristics. Reliability can also include estimates of a high degree of relationship across similar items within a single test or subtest that are intended to measure the same knowledge or skill. For judgmentally scored tests, such as essays, another widely used index of reliability addresses stability across raters or scorers. In each case, reliability can be estimated in different ways, using one of several statistical procedures.[95] Different kinds of reliability estimates vary in degree and nature of generalization. Readers are encouraged to review Chapter 2, Reliability and Errors of Measurement, in the *Joint Standards* for additional, relevant information.[96]

---

[90] *See* Standard 1.22, 1.23, 7.5 (n.71), 7.10 (n.33) and 13.9 (n.23) in *Joint Standards, supra* note 3, at pp. 23, 82, 83, 147.

Standard 1.22 states, "When it is clearly stated or implied that recommended test use will results in a specific outcome, the basis for expecing that outcome should be presented, together with relevant evidence." *Joint Standards, supra* note 3, at p. 23.

Standard 1.23 states, "When a test use or score interpretation is recommended on the grounds that testing or the testing program per se will result in some indirect benefit in addition to the utility of information from the test scores themselves, the rationale for anticipating the indirect benefit should be made explicit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. Due weight should be given to any contradictory findings in the scientific literature, including findings suggesting important indirect outcomes other than those predicted." *Joint Standards, supra* note 3, at p. 23.

[91] The Comment under Standard 13.1 states, "Mandated testing programs are often justified in terms of their potential benefits for teaching and learning. Concerns have been raised about the potential negative impact of mandated testing programs, particularly when they result directly in important decisions for individuals or institutions. Frequent concerns include narrowing the curriculum to focus only on the objectives tested, increasing the number of dropouts among students who do not pass the test, or encouraging other instructional or administrative practices simply designed to raise test scores rather than to affect the quality of education." *Joint Standards, supra* note 3, at p. 145.

[92] *High Stakes, supra* note 11, at pp. 247-272.

[93] Evaluating the reliability of test results includes identifying the major sources of measurement error, the size of the errors resulting from these sources, the indication of the degree of reliability to be expected, or the generalizability of results across items, forms, raters, sampling, administrations, and other measurement facets.

C. Fairness

Tests are fair when they yield score interpretations that are valid and reliable for all groups of students who take the tests. That is, the tests must measure the same academic constructs (knowledge and skills) for all groups of students who take them, regardless of race, national origin, gender, or disability. Similarly, it is important that the scores not substantially and

> **Fairness, like validity, cannot be properly addressed as an afterthought. ... It must be confronted throughout the interconnected phases of the testing process, from test design and development to administration, scoring, interpretation, and use.**
>
> National Research Council, ***High Stakes: Testing for Tracking, Promotion and Graduation***, pp. 80-81 (Jay P. Heubert & Robert M. Hauser eds., 1999).

---

[94] All sources of assessment information, including test results, include some degree of error. There are two types of error. The first is random error that affects scores in such a way that sometimes students will score lower and sometimes higher than their "true" score (the actual mastery level of the students' knowledge and skills). This type of error, also known as measurement error, particularly affects reliability of scores. Therefore, test scores are considered reliable when evidence demonstrates that there is a minimum amount of random measurement error in the test scores for a given group.

The second type of error that affects test results is systematic error. Systematic error consistently affects scores in one direction; that is, this type of error causes some students to consistently score lower or consistently score higher than their "true" (or actual) level of mastery. For instance, visually impaired students will consistently score lower than they should on a test which has not been administered for them in Braille or large print, because their difficulty in reading the items on the page will negatively impact their score. This type of error generally affects the validity of the interpretation of the test results and is discussed in the validity section above. Systematic error should also be minimized in a test for all test takers.

When educators and policy-makers are evaluating the adequacy of a test for their local population of students, it is important to consider evidence concerning both types of error.

[95] These types of reliability estimates are known as test-retest, alternate forms, internal consistency, and inter-rater estimates, respectively. *Joint Standards*, *supra* note 3, at pp. 25-31.

[96] *Joint Standards*, *supra* note 3, pp. 25-36.

systematically underestimate or overestimate the knowledge or skills of members of a particular group. The *Joint Standards* discusses fairness in testing in terms of lack of bias, equitable treatment in the testing process, equal scores for students who have equal standing on the tested constructs, and, depending on the purpose, equity in opportunity to learn the material being tested.[97] In order to promote readability, the discussion on fairness presented here is meant to reflect this complex topic in an accurate, but concise and user-friendly way. Readers are encouraged to review Chapter 7, Fairness in Testing and Test Use, in the *Joint Standards* for additional, relevant information.[98]

1. Fairness in Validity

Demonstrating fairness in the validation of test score inferences focuses primarily on making sure that the scores reflect the same intended knowledge and skills for all students taking the test. For the most part this means that the test should minimize the measurement of material that is extraneous to the intended constructs and that confounds the ability of the test to accurately measure the constructs that it intends to measure. A test score should accurately reflect how well each student has mastered the intended constructs. The score should not be significantly impacted by construct irrelevant influences.

The *Joint Standards* identifies a number of standards that outline important considerations related to fairness in validity throughout test development, test implementation, and the proper use of reported test results.[99]

Documenting fairness during test development involves gathering adequate evidence that items and test scores are constructed so that the inferences validly reflect what is intended. For all groups of test takers, evidence should support that valid inferences can be drawn from the scores.[100] The *Joint Standards* states that when credible research reports that item and test results differ in meaning across examinee subgroups, then, to the extent feasible, separate validity evidence should be collected for each relevant subgroup.[101] When items function differently across relevant subgroups, appropriate studies should be conducted, when feasible, so that bias in items due to test design, content, and format is detected and eliminated.[102] Developers should strive to identify and eliminate language, form, and content in tests that have a different meaning in one subgroup than in others, or that generally have sensitive connotations, except when judged to be necessary for adequate representation of the intended constructs.[103] Adequate subgroup analyses should be conducted when evaluating the validity of scores for prediction purposes.[104]

---

[97] *Joint Standards, supra* note 3, at pp. 74-80. In test measurement, the term fairness has a specific set of technical interpretations. Four of these interpretations are discussed in the *Joint Standards.* For instance, bias is discussed in relation to fairness and is defined in the *Joint Standards* in two ways: "In a statistical context, (bias refers to) a systematic error in a test score. In discussing test fairness, bias (also) may refer to construct underrepresentation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers." *Joint Standards, supra* note 3, at p. 172. Fairness as equitable treatment in the testing process "requires consideration not only of the test itself, but also the context and purpose of testing, and the manner for which test scores are used." *Joint Standards, supra* note 3, at p. 74. Equal scores for students of equal standing reflects that "examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership." *Joint Standards, supra* note 3, at p. 74. For purposes such as promotion and graduation, "[w]hen some test takers have not had the opportunity to learn the subject matter covered by the test content, they are likely to get low scores . . . low scores may have resulted in part from not having had the opportunity to learn the material tested as well as from having had the opportunity and failed to learn." *Joint Standards, supra* note 3, at p. 76.

[98] *Joint Standards, supra* note 3, at pp. 73-84.

[99] *Joint Standards, supra* note 3, at pp. 80-84.

[100] Standard 7.2 states, "When credible research reports differences in the effects of construct-irrelevant variance across subgroups of test takers on performance of some part of the test, the test should be used if at all only for those subgroups for which evidence indicates that valid inferences can be drawn from test scores." *Joint Standards, supra* note 3, at p. 81.

Adequate evidence should document the fair implementation of tests for all test takers. The testing process should reflect equitable treatment for all examinees.[105] The *Joint Standards* states, "In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct."[106]

---

[101] *See* Standard 7.1 and 7.3 in *Joint Standards, supra* note 3, at pp. 80-81.

Standard 7.1 states, "When credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup. Subgroups may be found to differ with respect to appropriateness of test content, internal structure of test responses, the relation of test scores to other variables, or the response processes employed by individual examinees. Any such findings should receive due consideration in the interpretation and use of scores as well as in subsequent test revisions." *Joint Standards, supra* note 3, at p. 80.

Standard 7.3 states, "When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups." *Joint Standards, supra* note 3, at p. 81.

[102] Standard 7.3 (n.101) in *Joint Standards, supra* note 3, at p. 81.

[103] *See* Standard 7.3 (n.101) and 7.4 in *Joint Standards, supra* note 3, at pp. 81-82.

Standard 7.4 states, "Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain." *Joint Standards, supra* note 3, at p. 82.

The Comment to Standard 7.4 states, "Two issues are involved. The first deals with the inadvertent use of language that, unknown to the test developer, has a different meaning or connotation in one subgroup than in others. Test publishers often conduct sensitivity reviews of all test material to detect and remove sensitive material from the test. The second deals with settings in which sensitive material is essential for validity. For example, history tests may appropriately include material on slavery or Nazis. Tests on subjects from life sciences may appropriately include material on evolution. A test of understanding of an organization's sexual harassment policy may require employees to evaluate examples of potentially offensive behavior." *Joint Standards, supra* note 3, at p. 82.

[104] *See* Standard 7.6 (n.89) in *Joint Standards, supra* note 3, at p. 82.

[105] Standard 7.12 states, "The testing or assessment process should be carried out so that test takers receive comparable and equitable treatment during all phases of the testing or assessment process." *Joint Standards, supra* note 3, at p. 84.

[106] Standard 7.7 in *Joint Standards, supra* note 3, at p. 82.

Documentation of appropriate reporting and test use should be available. Reported data should be clear and accurate, especially when there are high-stakes consequences for students.[107] When tests are used as part of decision-making that has high-stakes consequences for students, evidence of mean score differences between relevant subgroups should be examined, where feasible. When mean differences are found between subgroups, investigations should be undertaken to determine that such differences are not attributable to construct underrepresentation or construct irrelevant error.[108] Evidence about differences in mean scores and the significance of the validity errors should also be considered when deciding which test to use.[109] In using test results for purposes other than selection, a test taker's score should not be accepted as a reflection of standing on the intended constructs without consideration of alternative explanations for the test taker's performance.[110] Explanations might reflect limitations of the test, for instance construct irrelevant factors may have significantly impacted the student's score. Explanations may also reflect schooling factors external to the test, for instance lack of instructional opportunities.

The issue of feasibility in collecting validity evidence is discussed in a few of the standards summarized above. In the comments associated with these standards, feasibility is generally addressed in terms of adequate sample size, with continued operational use of a test as a way of accumulating adequate numbers of subgroup results over administrations. When credible research reports that results differ in meaning across subgroups, collecting separate and parallel types of validity data verifies that the same knowledge and skills are being measured for all groups of test takers. Particularly in high-stakes situations, it is important that all feasibility considerations include the potential costs to students of using information where the validity of the scores has not been verified.[111]

---

[107] *See* Standard 1.24 (n.89), 7.8, 7.9 and 7.10 (n.33) in *Joint Standards, supra* note 3, at pp. 23, 83.

Standard 7.8 states, "When scores are disaggregated and publicly reported for groups identified by characteristics such as gender, ethnicity, age, language proficiency, or disability, cautionary statements should be included whenever credible research reports that test scores may not have comparable meaning across these different groups." *Joint Standards, supra* note 3, at p. 83.

Standard 7.9 states, "When tests or assessments are proposed for use as instruments of social, educational, or public policy, the test developers or users proposing the test should fully and accurately inform policy-makers of the characteristics of the tests as well as any relevant and credible information that may be available concerning the likely consequences of test use." *Joint Standards, supra* note 3, at p. 83.

[108] Standard 7.10 (n.33) in *Joint Standards, supra* note 3, at p. 83.

[109] Standard 7.11 states, "When a construct can be measured in different ways that are approximately equal in their degree of construct representation and freedom from construct-irrelevant variance, evidence of mean score differences across relevant subgroups of examinees should be considered in deciding which test to use." *Joint Standards, supra* note 3, at p. 83.

[110] Standard 7.5 (n.71) in *Joint Standards, supra* note 3, at p. 82.

2. Fairness in Reliability

Fairness in reliability focuses on making sure that scores are stable and consistently accurate for all groups of students. Two key standards address this issue. First, when there are reasons for expecting that test reliability analyses might differ substantially for different subpopulations, reliability data should be presented as soon as feasible for each major population for whom the test is recommended.[112] Second, "[w]hen significant variations are permitted in test administration procedures, separate reliability analyses should be provided for scores produced under each major variation if adequate sample sizes are available."[113] Often, continued operational use of a test is a way to accumulate an adequate sample size over administrations.

D. Cut Scores

The same principles regarding validity, reliability, and fairness apply generally to the establishment and use of cut scores for the purpose of making high-stakes educational decisions. Cut scores, also known as cut points or cutoff scores, are specific points on the test or scale where test results are used to divide levels of knowledge, skill, or ability. Cut scores are used in a variety of contexts, including decisions for placement purposes or for other specific outcomes, such as graduation, promotion, or admissions.[114] A cut score may divide the demonstration of acceptable and unacceptable skills, as in placement in gifted and talented programs where students are accepted or rejected. There may be multiple cut scores that identify qualitatively distinct levels of performance. In order to promote readability, the discussion on cut scores presented here is meant to reflect this complex topic in an accurate, but concise and user-friendly way. Readers are encouraged to review Chapter 4, Scales, Norms, and Score Comparability, in the *Joint Standards*, for additional, relevant information about cut scores particularly pages 53-54.

---

[111] The Comment to Standard 10.7 states, "In addition to modifying tests and test administration procedures for people who have disabilities, evidence of validity for inferences drawn from these tests is needed. *Validation is the only way to amass knowledge about the usefulness of modified tests for people with disabilities. The costs of obtaining validity evidence should be considered in light of the consequences of not having usable information regarding the meanings of scores for people with disabilities.* This standard is feasible in the limited circumstances where a sufficient number of individuals with the same level or degree of a given disability is available." *Joint Standards, supra* note 3, at p. 107 (emphasis added).

[112] Standard 2.11 states, "If there are generally accepted theoretical or empirical reasons for expecting that reliability coefficients, standard errors of measurement, or test information functions will differ substantially for various subpopulations, publishers should provide reliability data as soon as feasible for each major population for which the test is recommended." *Joint Standards, supra* note 3, at p. 34.

It should be noted that reliability estimates may differ simply because of limited variance within a group. This is not a flaw in the test leading to unfairness, but rather a function of the statistical methodologies used in calculating the estimates.

[113] Standard 2.18 in *Joint Standards, supra* note 3, at p. 36.

[114] *See also* Standard 1.19 and 13.9 (n.23) in *Joint Standards, supra* note 3, at pp. 22, 147.

Standard 1.19 states, "If a test is recommended for use in assigning persons to alternative treatments or is likely to be so used, and if outcomes from those treatments can reasonably be compared on a common criterion, then, whenever feasible, supporting evidence of differential outcomes should be provided." *Joint Standards, supra* note 3, at p. 22.

> **Where the results of the [cutscore] setting process have highly significant consequences, ... those responsible for establishing cutscores should be concerned that the process...[is] clearly documented and defensible.**
>
> *Joint Standards*, Introduction to Chapter 4, p. 54.

Many of the concepts regarding test validity apply to cut scores—that is, the cut points themselves, like all scores, must be accurate representations of the knowledge and skills of students.[115] Further, "[w]hen feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria."[116] Validity evidence should generally be able to demonstrate that students above the cut score represent or demonstrate a qualitatively greater degree or different type of skills and knowledge than those below the cut score, whenever these types of inferences are made. In high-stakes situations, it is important to examine the validity of the inferences that underlie the specific decisions being made on the basis of the cut scores. In other words, what must be validated is the specific use of the test based on how the scores of students above and below the cut score are being interpreted.

Reliability of the cut scores is also important. The *Joint Standards* states that where cut scores are specified for selection or placement, the degree of measurement error around each cut score should be reported.[117] Evidence should also indicate the misclassification rates, or percentage of error in classifying students, that are likely to occur among students with comparable knowledge and skills.[118] This information should be available by group as soon

---

[115] *See Joint Standards, supra* note 3, pp. 9-16 (Chapter 1, Validity, discusses that the interpretation of *all* scores should be an accurate representation of what is being measured).

[116] Standard 4.20 in *Joint Standards, supra* note 3, at p. 60.

[117] Standard 2.14 states, "Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score." *Joint Standards, supra* note 3, at p. 35.

[118] "Where the purpose of measurement is classification, some measurement errors are more serious than others. An individual who is far above or far below the value established for pass/fail or for eligibility for a special program can be mismeasured without serious consequences. Mismeasurment of examinees whose true scores are close to the cut score is a more serious concern. . . . The term *classification consistency* or *inter-rater agreement*, rather than *reliability*, would be used in discussions of consistency of classification. Adoption of such usage would make it clear that the importance of an error of any given size depends on the proximity of the examinee's score to the cut score." *Joint Standards, supra* note 3, at p. 30.

BEST COPY AVAILABLE

as feasible if there is a prior probability that the misclassification rates may differ substantially by group.[119] Misclassification of students above or below the cut points can result in both false positive and false negative classifications.[120] As an example of false negative misclassifaction one might ask, what percentage of students who should be allowed to graduate would not be allowed to do so because of error due to the test rather than differences in their actual knowledge and skills? The *Joint Standards* states, "Adequate precision in regions of score scales where cut points are established is prerequisite to reliable classification of examinees into categories."[121]

There is no single right answer to the questions of when, where and how cut scores should be set on a test with high-stakes consequences for students.[122] Some experts suggest, however, that multiple standard-setting methods of determining cut scores should be used when determining a final cut score.[123] Further, the reasonableness of the standard setting process and the consequences for students should be clearly and specifically documented for a given use.[124] Both the *Joint Standards* and *High Stakes* repeatedly state that decisions should not be made solely or automatically on the basis of a single test score, and that other relevant information should be taken into account if it will enhance the overall validity of the decision.[125]

---

[119] Standard 2.11 (n.112) in *Joint Standards, supra* note 3, at p. 34.

[120] *Joint Standards, supra* note 3, at p. 30.

[121] *Joint Standards, supra* note 3, at p. 59.

[122] *High Stakes, supra* note 11, at p. 168.

[123] *High Stakes, supra* note 11, at p. 169.

[124] *See* Standard 4.19, 4.21 and their Comments in *Joint Standards, supra* note 3, at pp. 59-60; *see also High Stakes, supra* note 11, at pp. 89-187 (Chapters 5, 6, and 7).

Standard 4.19 states, "When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented." *Joint Standards, supra* note 3, at p. 59.

Standard 4.21 states, "When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way." *Joint Standards, supra* note 3, at p. 60.

[125] *See High Stakes, supra* note 11, at pp. 89-187 (Chapters 5, 6, and 7); Standard 13.7 (n.8) in *Joint Standards, supra* note 3, at p. 146.

Test Measurement Principles: Questions about Appropriate Test Use

In order to determine if a test is being used appropriately to make high-stakes decisions about students, considerations about the context of the test use need to be addressed, as well as the validity, reliability, and fairness of the score interpretations from the current test being proposed.

1. What is the purpose for which the test is being used?
2. What information, besides the test, is being collected to inform this purpose?
3. What are the particular propositions that need to be true to support the inferences drawn from the test scores for a given use?
4. Based on how the test results are to be used, is there adequate evidence of the propositions to document the validity of the inferences for students taking the test? For example:
   - Does the evidence support the proposition that the test measures the specific knowledge and skills the test developers say that it measures?
   - Does the evidence support the proposition that the interpretation of the test scores is valid for the stated purpose for which the test is being proposed?
   - Does the evidence support the proposition that the interpretation of the test scores is valid in the particular type of situation where the test is to be administered?
   - Does the evidence support the proposition that the interpretation of the test scores is valid for the specific groups of students who are taking the test?
5. Is there adequate evidence of reliability of the test scores for the proposed use?
6. Is there adequate evidence of fairness in validity and reliability to document that the test score inferences are accurate and meaningful for all groups of students taking the test? That is:
   - Does the evidence support the inference that the test is measuring the same constructs for all groups of students?
   - Does the evidence support that the scores do not systematically underestimate or overestimate the knowledge or skills of members of any particular group?
7. Is there adequate evidence that cutscores have been properly established and that they will be used in ways that will provide accurate and meaningful information for all test takers?

II. The Testing of All Students: Issues of Intervention and Inclusion

All aspects of validity, reliability, fairness, and cut scores discussed above are applicable to the measurement of knowledge and skills of all students, including limited English proficient students[126] and students with disabilities. This section addresses additional issues related to accurately measuring the knowledge and skills of these two populations in selected situations. Issues affecting limited English proficient and disabled students are addressed separately below following discussion of general considerations about the selection and use of accommodations.

Whenever tests are intended to evaluate the knowledge of skills of different groups of students, ensuring that test score inferences accurately reflect the intended constructs for all students is a complex task. It involves several aspects of test construction, pilot testing, implementation, analysis, and reporting. For limited English proficient students and students with disabilities, the appropriate inclusion of students from these groups in validation and norming samples, and the meaningful inclusion of limited English proficient and disability experts throughout the test development process, are necessary to ensure suitable test quality for these groups of test takers.

The proper inclusion of diverse groups of students in the same academic achievement testing program helps to ensure that high-stakes decisions are made on the basis of test results that are as comparable as possible across all groups of test takers.[127] If different tests are used as part of the testing program, it is important to ensure that they measure the same content standards. The appropriate inclusion of students can also help to ensure that educational benefits attributable to the high-stakes decisions will be available to all. In some cases, it is appropriate to test limited English proficient students and students with disabilities under standardized conditions, as long as the evidence supports the validity of the results in a given situation for these students. In other cases, the conditions may have to be accommodated to assure that the inferences of the scores validly reflect the students' mastery of the intended constructs.[128] The use of multiple measures generally enhances the accuracy of the educational decisions, and these measures can be used to confirm the validity of the test results. The use of multiple measures is particularly relevant for limited English proficient students and students with disabilities in cases where technical data are in the process of being collected on the proper use of accommodations and the proper interpretation of test results when testing conditions are accommodated.

---

[126] These are students who are learning English as a second language; the same population sometimes also is referred to as English language learners.

[127] See High Stakes, supra note 11, at pp. 7, 80.

[128] *See Joint Standards, supra* note 3 at pp. 71-80, 91-97, 101-106 (Chapters 7, 9, and 10).

A. General Considerations about Accommodations

Making similar inferences about scores from academic achievement tests for all test takers, and making appropriate decisions when using these scores, requires accurately measuring the same academic constructs (knowledge and skills in specific subject areas) across groups and contexts. In measuring the knowledge and skills of limited English proficient students and students with disabilities, it is particularly important that the tests actually measure the intended knowledge and skills and not factors that are extraneous to the intended construct.[129] For instance, impaired visual capacity may influence a student's test score in science when the student must sight read a typical paper and pencil science test. In measuring science skills, the student's sight likely is not relevant to the student's knowledge of science. Similarly, how well a limited English proficient student reads English may influence the student's test score in mathematics when the student must read the test. In this case, the student's reading skills likely are not relevant when the skills of mathematics computation are to be measured. The proper selection of accommodations for individual students and the determination of technical quality associated with accommodated test scores are complex and challenging issues that need to be addressed by educators, policy-makers, and test developers.

Standard 10.1

In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement.

Typically, accommodations to established conditions are found in three main phases of testing: 1) the administration of tests, 2) how students are allowed to respond to the items, and 3) the presentation of the tests (how the items are presented to the students on the test instrument). Administration accommodations involve setting and timing, and can include extended time to counteract the increased literacy demands for English language learners or fatigue for a student with sensory disabilities. Response accommodations allow students to demonstrate what they know in different ways, such as responding on a computer rather than in a test booklet. Presentation accommodations can include format variations such as fewer items per page, large print, and plain language editing procedures, which use short sentences, common words, and active voice. There is wide variation in the types of accommodations used across states and school districts. (Appendix C lists many of the accommodations used in large-scale testing for limited English proficient students and students with disabilities. The list is not meant to be exhaustive, and its use in this document should not be seen as an endorsement of any specific accommodations. Rather, the Appendix is meant to provide examples of the types of accommodations that are being used with limited English proficient students and students with disabilities.)

---

[129] This is known as construct irrelevance. *See* discussion *supra* Chapter 1 Part (I)(A)(3) (Sources of Validity Error); *Joint Standards*, *supra* note 3, at pp. 173-174.

Issues regarding the use of accommodations are complex. When the possible use of an accommodation for a student is being considered, two questions should be examined: 1) What is being measured if conditions are accommodated? 2) What is being measured if the conditions remain the same? The decision to use an accommodation or not should be grounded in the ultimate goal of collecting test information that accurately and fairly represents the knowledge and skills of the individual student on the intended constructs. The overarching concern should be that test score inferences accurately reflect the intended constructs rather than factors extraneous to the intent of the measurement.[130]

B. Testing of Limited English Proficient Students

The *Joint Standards* and several recent measurement publications discuss the population of limited English proficient students and how test publishers and users have handled inclusion in tests to date.[131] This section briefly outlines principles derived from the *Joint Standards* and these publications. It addresses two types of testing situations especially relevant for limited English proficient students: the assessment of English language proficiency and the assessment of academic educational achievement.

1. Assessing English Language Proficiency

Issues of validity, reliability, and fairness apply to tests and other relevant assessments that measure English language proficiency. English language proficiency is typically defined as proficiency in listening, speaking, reading, and writing English.[132] Assessments that measure English language proficiency are generally used to make decisions about who should receive English language acquisition services, the type of programs in which these students are placed, and the progress of students in the

Standard 9.10-

Inferences about test takers' general language proficiency should be based on tests that measure a range of language features, and not on a single linguistic skill.

---

[130] *See* Standard 9.1 and 10.1 in *Joint Standards, supra* note 3, at pp. 97, 106; Messick, *Validity, supra* note 85.

Standard 9.1 states, "Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences." *Joint Standards, supra* note 3, at p. 97.

Standard 10.1 states, "In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement." *Joint Standards, supra* note 3, at p. 106.

[131] *E.g., Joint Standards, supra* note 3, at pp. 91-97 (Chapter 9); *High Stakes, supra* note 11, at pp. 211-237 (Chapter 9); National Research Council, *Improving America's Schooling for Language Minority Children: A Research Agenda* (Diane August & Kenji Hakuta eds., 1997) (hereinafter *Improving America's Schooling for Language Minority Children*); Rebecca J. Kopriva, Council of Chief State School Officers, *Ensuring Accuracy in Testing for English Language Learners* (2000) (hereinafter Kopriva, *Ensuring Accuracy in Testing*).

[132] *Improving America's Schooling for Language Minority Children, supra* note 131, at pp. 116-118.

appropriate programs. They are also used to evaluate the English proficiency of students when exiting from a program or services, to ensure that they can successfully participate in the regular school curriculum. In making decisions about which tests are appropriate, it is particularly important to make sure that the tests accurately and completely reflect the intended English language proficiency constructs so that the students are not misclassified. It is generally accepted that an evaluation of a range of communicative abilities will typically need to be assessed when placement decisions are being made.[133]

2. Assessing the Academic Educational Achievement of Limited English Proficient Students

Several factors typically affect how well the educational achievement of limited English proficient students is measured on standardized academic achievement tests. Technical issues associated with developing meaningful achievement tests for limited English proficient students can be complex and challenging. For all test takers, any test that employs written or oral skills in English or in another language is, in part, a measure of those skills in the particular language. Test use with individuals who have not sufficiently acquired the literacy or fluency skills in the language of the test may introduce construct-irrelevant components to the testing process. Further, issues related to differences in the experiences of students may substantially affect how test items are interpreted by different groups of students. In both instances, test scores may not accurately reflect the qualities and competencies that the test intends to measure.[134]

a. Background Factors for Limited English Proficient Students

The background factors particularly salient in ensuring accuracy in testing for students with limited English proficiency tend to relate to language proficiency, culture, and schooling.[135]

Limited English proficient students often bring varying levels of English and home-language fluency and literacy skills to the testing situation. These students may be adept in conversing orally in their home language, but unless they have had formal schooling in their home language, they may not have a corresponding level of literacy. Also, while students with limited English proficiency may acquire a degree of fluency in English, literacy in English for many students comes later. To add to the complexity, proficiency in fluency and literacy in either the home language or English involves both social and academic components. Thus, a student may be able to write a well-organized social letter in his or

---

[133] Standard 9.10 and Comment in *Joint Standards, supra* note 3, at pp. 99-100.

Standard 9.10 states, "Inferences about test takers' general language proficiency should be based on tests that measure a range of language features, and not on a single linguistic skill." *Joint Standards, supra* note 3, at pp. 99-100.

[134] *Joint Standards, supra* note 3, at pp. 91-97.

[135] *See Joint Standards, supra* note 3, at pp. 91-100 (Chapter 9); *Improving Schooling for Language Minority Children, supra* note 131; Kopriva, *Ensuring Accuracy in Testing, supra* note 131, at pp. 9-11 (Introduction).

Factors Related to Accurately Testing Limited English Proficient Students

Language Proficiency
- The student's level of oral and written proficiency in English
- The student's proficiency in his or her home language
- The language of instruction

Cultural Issues
- Background experiences
- Perceptions of prior experiences
- Value systems,

Schooling Issues
- The amount of formal elementary and secondary schooling in the student's home country, if applicable, and in U.S. schools
- Consistency of schooling
- Instructional practices in the classroom

her home language, and may not be able to orally explain adequately in that language how to solve a mathematics problem that includes the knowledge of concepts and words endemic to the field of mathematics. The same phenomena may occur in English as well.[136]

Therefore, in determining how to effectively measure the academic knowledge and skills of limited English proficient students, educators and policy-makers should consider how to minimize the influence of literacy issues, except when these constructs are explicitly being measured. The levels of proficiency of limited English proficient students in their home language and in English, as well as the language of instruction, are important in determining in which language an achievement test should be administered, and which accommodations to standardized testing conditions, if any, might be most useful for which students.[137]

Additionally, diverse cultural and other background experiences, including variations in amount, type and location (home country and United States) of formal elementary and secondary schooling, as well as interrupted and multi-location schooling of students (of the type frequently experienced by children of migrant workers), affect language literacy, the contextual content of items, and the academic foundational knowledge base that can be assumed in appropriately interpreting the results of educational achievement tests. The format and procedures involved in testing can also affect accuracy in test scores, particularly if the test practices differ substantially from ongoing instructional practices in classrooms, including which accommodations are used in the classroom and how they are used.[138]

---

[136] *Improving America's Schooling for Language Minority Children, supra* note 131, at pp. 113-137.

[137] *Improving America's Schooling for Language Minority Children, supra* note 131, at pp. 113-137.

[138] Kopriva, *Ensuring Accuracy in Testing, supra* note 131, at pp. 29-48, 61-70, 95-98.

b. Including Limited English Proficient Students in Large-Scale Standardized Achievement Tests

The *Joint Standards* recognizes the complexity of developing educational achievement tests that are appropriate for a range of test takers, including those who are limited English proficient. Overall, "testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences."[139] When credible research evidence reports that scores may differ in meaning across subgroups of linguistically diverse test takers, then, to the extent feasible, the same form of validity evidence should be collected for each relevant subgroup as for the examinee population as a whole.[140] The *Joint Standards* states, "When a test is recommended for use with linguistically diverse test takers, test developers and publishers should provide the information necessary for appropriate test use and interpretation."[141] Furthermore, "when testing an examinee proficient in two or more languages for which the test is available, the examinee's relative language proficiencies should be determined. The test generally should be administered in the test taker's most proficient language, unless proficiency in the less proficient language is part of the assessment."[142] Recommended accommodations should be used appropriately and described in detail in the test manual;[143] translation methods and interpreter expertise should be clearly described;[144] evidence of test comparability should be reported when multiple language versions of a test are intended to be comparable;[145] and evidence of the score reliability and the validity of the translated test's score inferences should be provided for the intended uses and linguistic groups.[146]

---

[139] Standard 9.1 in *Joint Standards, supra* note 3, at p. 97.

[140] Standard 9.2 states, "When credible research evidence reports that test scores differ in meaning across subgroups of linguistically diverse test takers, then to the extent feasible, test developers should collect for each linguistic subgroup studied the same form of validity evidence collected for the examinee population as a whole." *Joint Standards, supra* note 3, at p. 97.

[141] Standard 9.6 in *Joint Standards, supra* note 3, at p. 99.

[142] Standard 9.3 in *Joint Standards, supra* note 3, at p. 98.

[143] See Standard 9.4 and 9.5 in *Joint Standards, supra* note 3, at p. 98.

Standard 9.4 states, "Linguistic modifications recommended by test publishers, as well as the rationale for the modifications, should be described in detail in the test manual." *Joint Standards, supra* note 3, at p. 98.

Standard 9.5 states, "When there is credible evidence of score comparability across regular and modified tests or administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores." *Joint Standards, supra* note 3, at p. 98.

[144] See Standard 9.7 and 9.11 in *Joint Standards, supra* note 3, at pp. 99-100.

Standard 9.7 states, "When a test is translated from one language to another, the methods used in establishing the adequacy of the translation should be described, and empirical and logical evidence should be provided for score reliability and the validity of the translated test's score inferences for the uses intended in the linguistic groups to be tested." *Joint Standards, supra* note 3, at p. 99.

Standard 9.11 states, "When an interpretation is used in testing, the interpreter should be fluent in both the language of the test and the examinee's native language, should have expertise in translating, and should have a basic understanding of the assessment process." *Joint Standards, supra* note 3, at p. 100.

Providing accommodations to established testing conditions for some students with limited English proficiency may be appropriate when their use would yield the most valid scores on the intended academic achievement constructs. Deciding which accommodations to use for which students usually involves an understanding of which construct irrelevant background factors would substantially influence the measurement of intended knowledge and skills for individual students, and if the accommodations would enhance the validity of the test score interpretations for these students.[147] In collecting evidence to support the technical quality of a test for limited English proficient students, the accumulation of data may need to occur over several test administrations to ensure sufficient sample sizes. Educators and policy-makers need to understand that the proper use of accommodations for limited English proficient students and the determination of technical quality are complex and challenging endeavors.

Appendix C lists various test presentation, administration, and response accommodations that states and districts generally employ when testing limited English proficient students. Examples of accommodations in the presentation of the test include editing text so the items are in plain language, or providing page formats which minimize confusion by limiting use of columns and the number of items per page. Presenting the test in the student's native language is an accommodation to a test written in English when the same constructs are being measured on both the English- and native-language versions. It is essential that translations accurately convey the meaning of the test items; poor translations can prove more harmful than helpful.[148] Administration accommodations include extending the length of the testing period, permitting breaks, administering tests in small groups or in separate rooms, and allowing English or native-language glossaries or dictionaries as appropriate. Response accommodations include oral response and permitting students to respond in their native language.

---

[145] Standard 9.9 states "When multiple language versions of a test are intended to be comparable, test developers should report evidence of test comparability." *Joint Standards, supra* note 3, at p. 99.

[146] Standard 9.7 (n.144) and Comment in *Joint Standards, supra* note 3, at p. 99.

The Comment to Standard 9.7 states "[f]or example, if a test is translated into Spanish for use with Mexican, Puerto Rican, Cuban, Central American, and Spanish populations, score reliability and the validity of the test score inferences should be established with members of each of these groups separately where feasible. In addition, the test translation methods used need to be described in detail." *Joint Standards, supra* note 3, at p. 99.

[147] Kopriva, *Ensuring Accuracy in Testing, supra* note 131, at pp. 49-66, 71-76 (discussing which accommodations might be most beneficial for students with various background factors).

[148] President's Advisory Commission on Educational Excellence for Hispanic Americans, *Testing Hispanic Students in the United States: Technical and Policy Issues, Executive Summary*, p. 8 (2000).

C. Testing of Students with Disabilities

The *Joint Standards* and several recent measurement publications discuss the population of students with disabilities and how test publishers and users have handled inclusion in tests to date.[149] This section briefly outlines principles derived from the *Joint Standards* and these publications. It addresses three types of testing situations especially relevant for students with disabilities: tests used for diagnostic and intervention purposes, the assessment of academic educational achievement, and alternate assessments for elementary and secondary school students with disabilities who cannot participate in districtwide academic achievement tests.

1. Tests Used for Diagnostic and Intervention Purposes

All issues of validity, reliability, and fairness apply to tests and other assessments used to make diagnostic and intervention decisions for students with disabilities. Tests that yield diagnostic information typically focus in great detail on identifying the specific challenges and strengths of a student.[150] These diagnostic tests are often administered in one-to-one situations (test taker and examiner) rather than in a group situation. In many cases, they have been designed with standardized adaptations to fit the needs of individual examinees. In making decisions about which tests are appropriate to use, it is important to make sure that the tests accurately and completely reflect the intended constructs, so that the interventions are appropriate and beneficial for the individual students. Proper analyses should be conducted to yield correct interpretations of results when differential prediction for different groups is likely.[151]

Standard 10.12

In testing individuals with disabilities for diagnostic and intervention purposes, the test should not be used as the sole indicator of the test taker's functioning. Instead, multiple sources of information should be used.

2. Assessing the Academic Educational Achievement of Students with Disabilities

Several factors affect how well the educational achievement of students with disabilities is measured on standardized academic achievement tests. Test scores should accurately measure the students' knowledge and skills in academic achievement rather than factors irrelevant to the intended constructs of the test.[152] The technical issues associated with developing meaningful achievement tests for students with disabilities can be complex and challenging. Under federal law, students with disabilities must be included in statewide or districtwide assessment programs and provided with appropriate accommodations if necessary. Guidance about testing elementary and secondary school students with disabilities is addressed by the individualized education program (IEP) process or other applicable evaluation procedures. The IEP or Section 504 plan addresses how a student should be tested, and identifies testing accommodations that would be appropriate for the individual student. The Individuals with Disabilities Education Act (IDEA) also requires state or local education agencies to develop guidelines for the relatively small number of students with disabilities who cannot take part in statewide or districtwide tests to participate in alternate assessments. The *Joint Standards* emphasizes that people who make decisions about accommodations for students with disabilities should be knowledgeable about the effects of the disabilities on test performance.[153]

---

[149] *E.g.*, *Joint Standards*, *supra* note 3, at pp. 101-106 (Chapter 10); *High Stakes*, *supra* note 11, at pp. 188-210 (Chapter 8); National Research Council, *Educating One and All: Students with Disabilities and Standards-Based Reform* (Lorraine M. McDonnell, Margaret J. McLaughlin & Patricia Morison eds., 1997) (hereinafter *Educating One and All*); Martha Thurlow, Judy Elliott & Jim Ysseldyke, *Testing Students with Disabilities* (1998) (hereinafter Thurlow et al., *Testing Students with Disabilities*).

[150] *Joint Standards*, *supra* note 3, at pp. 101-106, 119-145 (Chapters 10, 12, and 13); *High Stakes*, *supra* note 11, at pp. 13-28 (Chapter 1).

[151] *See* Standard 7.6 (n.89) in *Joint Standards*, *supra* note 3, at p. 82.

a. Background Factors for Students with Disabilities

The background factors particularly important to students with disabilities are generally related to the nature of the disabilities or to the schooling experiences of these students.[154] Within any disability category, the type, number, and severity of impairments vary greatly.[155] For instance, some students with learning disabilities have a processing disability in only one subject, such as mathematics, while others experience accessing, retrieving, and processing impairments that affect a broad number of school subjects and contexts. For many of these students, one or more of the impairments may be relatively mild, while for others one or more can be significant. Further, different types of disabilities yield significantly different constellations of issues. For instance, the considerations surrounding students with hearing impairments or deafness may overlap significantly with limited English proficient students in some ways and with other students with disabilities in other respects. The *Joint Standards* discusses provisions regarding the testing and validation of tests for limited English proficient students that apply to students who have hearing impairments or deafness, as well.[156] This complexity poses a challenge not only to educators, but also to test administrators and developers. In general, in determining how to use academic tests appropriately for students with disabilities, educators and policy-makers should consider how to minimize the influence of the impairments in measuring the intended constructs.

---

[152] Standard 10.1 (n.130) in *Joint Standards, supra* note 3, at p. 106.

[153] Standard 10.2 states, " People who make decisions about accommodations and test modification for individuals with disabilities should be knowledgeable of existing research on the effects of the disabilities in question on test performance. Those who modify tests should also have access to psychometric expertise for so doing. *Joint Standards, supra* note 3, at p. 106.

[154] *See Joint Standards, supra* note 3, at pp. 101-108 (Chapter 10); *Educating One and All, supra* note 149.

[155] Thurlow et al., *Testing Students with Disabilities, supra* note 149.

[156] *See* Standard 9.2 (n.140) and 9.10 (n.133) in *Joint Standards, supra* note 3, at pp. 97, 99-100.

Educating One and All explains that the schooling experiences of students with disabilities vary greatly as a function of their disability, the severity of impairments, and expectations of their capabilities.[157] Two sets of educational experiences, in particular, affect how educators and policy-makers accommodate tests and use them appropriately for this population. First, the IEP teams identify individual educational plans for students with disabilities that have different degrees of overlap with the general education curricula.

This alignment will affect what opportunities students with disabilities will have to master the material being tested on the schoolwide academic achievement tests. Second, the IEP team also recommends appropriate accommodations for students, and these accommodations are usually consistent with classroom accommodation techniques. However, while special educators have a long history of accommodating instruction and evaluation to fit student strengths, not all the instructional or testing practices in the classroom are appropriate in large-scale testing. Additionally, some students may not have been exposed routinely to the types of accommodations that would be possible in large-scale testing.[158]

Factors Related to Accurately Testing Students with Disabilities

Disability Issues
- Types of impairments
- Severity of impairments

Schooling Experiences
- Overlap of individualized educational goals and general education curricula in elementary and secondary schooling
- Pace of schooling
- Instructional practices in the classroom

b. Including Students with Disabilities in Large-Scale Standardized Achievement Tests

The *Joint Standards* recognizes the complexity of developing educational achievement tests that are appropriate for a range of test takers, including students with disabilities. The interpretation of the scores of students with disabilities should accurately and fairly reflect the academic knowledge, skills, or abilities that the test intends to measure. The interpretation should not be confounded by those challenges students face that are extraneous to the intent of the measurement.[159] Rather, validity evidence should document that the inferences of the scores of students with disabilities are accurate. Pilot testing and other technical investigations should be conducted where feasible to ensure the validity of the test inferences when accommodations have been allowed.[160] While, feasibility is a consideration, the *Joint Standards* comments that "the costs of obtaining validity evidence should be considered in light of the consequences of not having usable information regarding the meanings of scores for people with disabilities."[161]

---

[157] *Educating One and All, supra* note 149, at Chapter 3.

[158] *Educating One and All, supra* note 149, at Chapter 5.

---

[159] *See* Standard 10.1 (n.130) and 10.10 in *Joint Standards, supra* note 3, at pp. 106, 107-108.

Standard 10.10 states, "Any test modifications adopted should be appropriate for the individual test taker, while maintaining all feasible standardized features. A test professional needs to consider reasonably available information about each test taker's experiences, characteristics, and capabilities that might impact test performance, and document the grounds for the modification." *Joint Standards, supra* note 3, at pp. 107-108.

[160] Several standards discuss the appropriate types of validity evidence, including Standards 10.3, 10.5, 10.6, 10.7, 10.8, and 10.11. Because of the low-incidence nature of several of the disability groups, such as hearing loss, vision loss, or concomitant hearing and vision loss, especially when different severity levels and combinations of impairments are considered, this type of evidence will probably need to be accumulated over time in order to have a large enough sample size.

Standard 10.3 states, "Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications." *Joint Standards, supra* note 3, at p. 106.

Standard 10.5 states, "Technical material and manuals that accompany modified tests should include a careful statement of the steps taken to modify the test to alert users to changes that are likely to alter the validity of inferences drawn from the test scores." *Joint Standards, supra* note 3, at p. 106.

Standard 10.6 states, "If a test developer recommends specific time limits for people with disabilities, empirical procedures should be used, whenever possible, to establish time limits for modified forms of timed tests rather than simply allowing test takers with disabilities a multiple of the standard time. When possible, fatigue should be investigated as a potentially important factor when time limits are extended." *Joint Standards, supra* note 3, at p. 107.

Standard 10.7 states, "When sample sizes permit, the validity of inferences made from test scores and the reliability of scores on tests administered to individuals with various disabilities should be investigated and reported by the agency or publisher that makes the modification. Such investigations should examine the effects of modifications made for people with various disabilities on resulting scores, as well as the effects of administering standard unmodified tests to them." *Joint Standards, supra* note 3, at p. 107.

Standard 10.8 states, "Those responsible for decisions about test use with potential test takers who may need or may request specific accommodations should (a) possess the information necessary to make an appropriate selection of measures, (b) have current information regarding the availability of modified forms of the test in question, (c) inform individuals, when appropriate, about the existence of modified forms, and (d) make these forms available to test takers when appropriate and feasible." *Joint Standards, supra* note 3, at p. 107.

Standard 10.11 states, "When there is credible evidence of score comparability across regular and modified administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores." *Joint Standards, supra* note 3, at p. 108.

[161] *See* Comment to Standard 10.7 (n.111) in *Joint Standards, supra* note 3, at p. 106.

Providing accommodations to established testing conditions for some students with disabilities may be appropriate when their use would yield the most valid scores on the intended academic achievement constructs. Deciding which accommodations to use for which students usually involves an understanding of which construct irrelevant background factors would substantially influence the measurement of intended knowledge and skills for individual students, and if the accommodations would enhance the validity of the test score interpretations for these students.[162] In collecting evidence to support the technical quality of the test results for students with disabilities, the accumulation of data may need to occur over several administrations to ensure sufficient sample sizes. Educators and policy-makers need to understand that the proper use of accommodations for students with disabilities and the determination of technical quality are complex and challenging endeavors.

Appendix C lists various presentation, administration, and response accommodations that states and districts generally employ when testing students with disabilities. Examples of presentation accommodations are the use of Braille, large print, oral reading, or providing page formats that minimize confusion by limiting use of columns and the number of items per page. Administration accommodations in setting include allowing students to take the test at home or in a small group, and accommodations in timing include extended time and frequent breaks. Variations in response formats include allowing students to respond orally, point, or use a computer.

3. Alternate Assessments

Alternate assessments are assessments for those elementary and secondary school students with disabilities who cannot participate in state or districtwide standardized assessments, even with the use of appropriate accommodations and modifications.[163] For the constructs being measured, the considerations with respect to validity, reliability, and fairness apply to alternate assessments, as well. Appropriate content needs to be identified, and procedures need to be designed to ensure technical rigor.[164] In addition, evidence should show that the test measures the knowledge and skills it intends to measure, and that the measurement is a valid reflection of mastery in a range of contextual situations.

---

[162] Thurlow et al., *Testing Students with Disabilities, supra* note 149, for a discussion of which accommodations might be most beneficial for students with various impairments and other background factors.

[163] The IDEA requires use of alternate assessments in certain areas. See 34 C.F.R. § 300.138. These assessments may or may not be used in decisions that have high-stakes consequences for students.

[164] See *Educating One and All, supra* note 149, at Chapter 5, and Thurlow et al., *Testing Students with Disabilities, supra* note 149, for a discussion of the issues and processes involved in developing and implementing alternate assessments.

CHAPTER 2: Legal Principles

It is important for educators and policy-makers to understand the test measurement principles and the legal principles that will enable them to ask informed questions and make sound decisions regarding the use of tests for high-stakes purposes. The goal of this chapter is to explain the legal principles that apply to educational testing.

The primary focus of this chapter is four federal nondiscrimination laws, enacted by Congress, and their implementing regulations: Title VI of the Civil Rights Act of 1964 (Title VI), Title IX of the Education Amendments of 1972 (Title IX), Section 504 of the Rehabilitation Act of 1973 (Section 504), and Title II of the Americans with Disabilities Act of 1990 (Title II).[165] Within the U.S. Department of Education, the Office for Civil Rights has responsibility for enforcing the requirements of these four statutes and their implementing regulations. Although the Office for Civil Rights does not enforce federal constitutional provisions, an overview of these constitutional principles, including under the Fifth and Fourteenth Amendments of the U.S. Constitution, has also been included for informational purposes because of their importance to sound test use. The discussion of legal principles in this chapter is intended to reflect existing legal principles and does not establish new requirements.[166]

Some of the issues that have been considered by federal courts in assessing the legality of specific testing practices for making high-stakes decisions include:

- **The use of an educational test for a purpose for which the test was not designed or validated;[167]**
- **The use of a test score as the sole criterion for the educational decision;[168]**
- **The nature and quality of the opportunity provided to students to master required content, including whether classroom instruction includes the material covered by a test administered to determine student achievement;[169]**
- **The significance of any fairness problems identified, including evidence of differential prediction of a criterion and possible cultural biases in the test or in test items;[170] and**
- **The educational basis for establishing passing or cut-off scores.[171]**

---

[165] Title VI prohibits discrimination on the basis of race, color and national origin by recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Title VI is found at 34 C.F.R. Part 100. Title IX prohibits discrimination on the basis of sex by recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Title IX is found at 34 C.F.R. Part 106. Section 504 prohibits discrimination on the basis of disability by recipients of federal financial assistance. The U.S. Department of Education's regulation implementing Section 504 is found at 34 C.F.R. Part 104. Title II prohibits discrimination on the basis of disability by public entities, regardless of whether they receive federal funding. The U.S. Department of Justice's regulation implementing Title II is found at 28 C.F.R. Part 35.

[166] Consistent with this approach, court decisions are not cited if the case is still on appeal or the time to request an appeal has not ended.

I. Discrimination Under Federal Statutes and Regulations

Congress has enacted four statutes prohibiting discrimination based on race, color, national origin, sex, and disability in elementary and secondary schools, colleges, and universities. Title VI prohibits discrimination based on race, color, or national origin; Title IX prohibits discrimination based on sex; and Section 504 and Title II of the Americans with Disabilities Act (ADA) prohibit discrimination based on disability. Title VI, Title IX, and Section 504 apply to all educational institutions that receive federal funds. Title II of the ADA applies

---

[167] *See Sharif v. New York State Educ. Dep't.*, 709 F. Supp. 345, 354-55, 364 (S.D.N.Y. 1989) (in granting a motion for preliminary injunction, where girls received comparatively lower scores than boys, court found that the state's use of SAT scores as the sole basis for decisions awarding college scholarships intended to reward high school achievement was not educationally justified for this purpose in that the SAT had been designed as an aptitude test to predict college success and was not designed or validated to measure past high school achievement).

[168] *See id.* at 364; *see also United States v. Fordice*, 505 U.S. 717, 735-39 (1992) (holding that the state's reliance on minimum ACT scores was constitutionally suspect where the ACT requirement was originally adopted for discriminatory purposes, the current requirement was traceable to that decision and continued to have segregative effects, and the state failed to show that the "ACT-only" admissions standard was not susceptible to elimination without eroding sound educational policy, and recognizing that "[a]nother constitutionally problematic aspect of the state's use of the ACT test scores is its policy of denying automatic admission if an applicant fails to earn the minimum ACT score specified for the particular institution, without also resorting to the applicant's high school grades as an additional factor in predicting college performance."); *GI Forum Image De Tejas v. Texas Education Agency*, 87 F. Supp. 2d 667 (W.D. Tex. 2000) (upholding the use of Texas Assessment of Academic Skills examination as a requirement for high school graduation where the court found that the test was strongly correlated to the material actually taught in the classroom; minority students received an equal opportunity to learn the items presented on the test; the test had been extensively validated as a tool for measuring legislatively established minimum skills as a requisite for graduation; and multiple opportunities were provided to each student to pass the examination in conjunction with state mandated remediation targeted to the student's deficiency areas).

[169] *See Lau v. Nichols*, 414 U.S. 563, 566-69 (1974) (finding a violation of the Title VI regulations where limited English proficient students were taught only in English and not provided any special assistance needed to meet English language proficiency standards required by the state for a high school diploma); *see also Debra P. v. Turlington*, 644 F.2d 397, 406-08 (5th Cir. 1981) (holding that use of a graduation test that covered material that had not been taught in class would violate the due process and equal protection clauses and that, under the circumstances of the case, immediate use of the diploma sanction for test failure would punish black students for deficiencies created by an illegally segregated school system which had provided them with inferior physical structures, course offerings, instructional materials, and equipment).

[170] *See Larry P. v. Riles*, 793 F.2d 969, 980-81, 983 (9th Cir. 1984) (finding that IQ tests the state used had not been validated for use as the sole means for determining that black children should be placed in classes for educable mentally retarded students); *Sharif*, 709 F. Supp. at 354 (observing that the SAT under-predicts success for female college freshmen as compared with males); *see also Parents in Action on Special Educ. v. Hannon*, 506 F. Supp. 831, 836-37 (N.D. Ill. 1980) (court's analysis of items on I.Q. test found only minimal amount of cultural bias not resulting in erroneous mental retardation diagnoses given other information considered in process).

[171] *See Groves v. Alabama State Bd. of Educ.*, 776 F. Supp. 1518, 1530-31 (M.D. Ala. 1991) (finding test required for admission to undergraduate teacher training program would not be educationally justified if the passing score is not itself a valid measure of the minimal ability necessary to become a teacher); *Richardson v. Lamar County Bd. of Educ.*, 729 F. Supp. 806, 823-25 (M.D. Ala. 1989) (evidence revealed that cut-off scores had not been set through a well-conceived, systematic process nor could the scores be characterized as reflecting the good faith exercise of professional judgment), *aff'd sub nom.*, *Richardson v. Alabama State Bd. of Educ.*, 935 F.2d 1240 (11th Cir. 1991).

BEST COPY AVAILABLE

to public entities, including public school districts and state colleges and universities.[172] The Title VI, Title IX, Section 504, and Title II statutes and their implementing regulations as well as the equal protection clause of the Fourteenth Amendment to the United States Constitution, prohibit intentional discrimination, based on race, national origin, sex, or disability.[173] In addition, the regulations that implement Title VI, Title IX, Section 504 and Title II prohibit policies or practices that have a discriminatory disparate impact on students based on their race, national origin, sex, or disability.[174]

This section describes two central analytical frameworks for examining allegations of discrimination as set forth in federal nondiscrimination regulations: different treatment and disparate impact.[175] It also includes a further discussion of legal principles that apply specifically to students with limited English proficiency and to students with disabilities.

---

[172] OCR enforces five nondiscrimination statutes, Title VI of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000d*et seq.* (2000); Title IX of the Education Amendments of 1972, 20 U.S.C. §§ 1681 *et seq.* (1999); Section 504 of the Rehabilitation Act of 1973, as amended, 29 U.S.C. § 794 (1999); Title II of the Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12131 *et seq.* (1995 & Supp. 1999); and the Age Discrimination Act of 1975, as amended, 42 U.S.C. §§ 6101 *et. seq.* (1995 & Supp. 1999). Regulations issued by the United States Department of Education implementing Title VI, Title IX, and Section 504, respectively, can be found at 34 C.F.R. Part 100, 34 C.F.R. Part 106, and 34 C.F.R. Part 104. These regulations can be found on OCR's web site at *www.ed.gov/offices/OCR.* Regulations implementing Title II of the ADA can be found at 28 C.F.R. Part 35. Title III of the ADA, which is enforced by the U.S. Department of Justice, prohibits discrimination in public accommodations by private entities, including schools. Religious entities operated by religious organizations are exempt from Title III.

[173] The United States Supreme Court has held that "Title VI itself directly reached only instances of intentional discrimination . . . [but that] actions having an unjustifiable disparate impact on minorities could be addressed through agency regulations designed to implement the purposes of Title VI." *Alexander v. Choate*, 439 U.S. 287, 295 (1985), discussing *Guardians Ass'n v. City Service Comm'n of N.Y.*, 403 U.S. 582 (1983). The United States Supreme Court has never expressly ruled on whether Section 504, Title II and Title IX statutes prohibit not only intentional discrimination, but, unlike Title VI, prohibit disparate impact discrimination as well. *See, e.g., Choate*, 409 U.S. at 294-97 & n.11 (observing that Congress might have intended the Section 504 statute itself to prohibit disparate impact discrimination). Section 504 and Title II require reasonable modifications where necessary to enable persons with disabilities to participate in or enjoy the benefits of public services. Regardless, the regulations implementing Section 504, Title II, and Title IX, like the Title VI regulation, explicitly prohibit actions having discriminatory effects as well as actions that are intentionally discriminatory.

[174] 34 C.F.R. § 100.3(b)(2) (Title VI); 34 C.F.R. §§ 106.21(b)(2), 106.36(b), 106.52 (Title IX); 34 C.F.R. § 104.4(b)(4)(i) (Section 504); 28 C.F.R. § 35.130(b)(3) (Title II).

The authority of federal agencies to issue regulations with an "effects" standard has been consistently acknowledged by United States Supreme Court decisions and applied by lower federal courts addressing claims of discrimination in education. *See, e.g., Choate*, 469 U.S. at 289-300 (1985); *Guardians Ass'n*, 463 U.S. at 584-93; *Lau*, 414 U.S. at 568; *see also* Memorandum from the Attorney General for Heads of Departments and Agencies that Provide Federal Financial Assistance, *Use of the Disparate Impact Standard in Administrative Regulations under Title VI of the Civil Rights Act of 1964* (July 14, 1994).

[175] Intentional racial discrimination is a violation of both the Fourteenth Amendment to the United States Constitution and federal civil rights statutes in cases where evidence demonstrates that an action such as the use of a test for high-stakes purposes is motivated by an intent to discriminate. *See Elston v. Talladega County Bd. of Educ.*, 997 F.2d 1394, 1406 (11th Cir. 1993). As explained further in this section, the regulations promulgated under the federal civil rights statutes prohibit the use of neutral criteria having disparate effects unless the criteria are educationally justified. *See Guardians Ass'n*, 463 U.S. at 598.

A. Different Treatment

Under federal law, policies and practices generally must be applied consistently to similarly situated individuals or groups, regardless of their race, national origin, sex, or disability.[176] For example, a federal court concluded that a school district had intentionally treated students differently on the basis of race where minority students whose test scores qualified them for two or more ability levels were more likely to be assigned to the lower-level class than similarly situated white students, and no explanatory reason was evident.[177]

In addition, educational systems that previously discriminated by race in violation of the Fourteenth Amendment and have not achieved unitary status have an obligation to dismantle their prior *de jure* segregation. In such instances, school districts are under "a 'heavy burden' of showing that actions that [have] increased or continued the effects of the dual system serve important and legitimate ends."[178] When such a school district or other educational system uses a test or assessment procedure for a high-stakes purpose that has significant racially disparate effects, to justify the test use, the school district must show that the test results are not due to the present effects of prior segregation or that the practice or procedure remedies the present effects of such segregation by offering better educational opportunities.[179]

---

[176] For example, under the Fourteenth Amendment and Title VI, different treatment based on race or ethnicty is permitted only when such action is narrowly tailored to further a compelling state interest. *See Adarand Constructors, Inc., v. Pena*, 515 U.S. 200 (1995); *Richmond v. Croson*, 488 U.S. 469 (1989); *Regents of the Univ. of Cal. v. Bakke*, 438 U.S. 265 (1978).

[177] *People Who Care v. Rockford Bd. of Educ.*, 851 F. Supp. 905, 958-1001 (N.D. Ill. 1994), *remedial order rev'd. in part*, 111 F.3d 528 (7th Cir. 1997). On appeal, the Seventh Circuit Court of Appeals stated that the appropriate remedy based on the facts in the case was to require the district to use objective, non-racial criteria to assign students to classes, rather than abolishing the district's tracking system. *People Who Care*, 111 F.3d at 536.

[178] *Dayton Bd. of Educ. v. Brinkman*, 443 U.S. 526, 538 (1979) (quoting *Green v. County School Bd.*, 391 U.S. 430, 439 (1968)).

[179] *See Debra P. v. Turlington*, 644 F.2d 397, 407 (5th Cir. 1981) ("[Defendants] failed to demonstrate either that the disproportionate failure [rate] of blacks was not due to the present effects of past intentional segregation or, that as presently used, the diploma section was necessary [in order] to remedy those effects."); *McNeal v. Tate County Sch. Dist.*, 508 F.2d 1017, 1020 (5th Cir. 1975) (ability grouping method that causes segregation may nonetheless be used "if the school district can demonstrate that its assignment method is not based on the present results of past segregation or that the method of assignment will remedy such effects through better educational opportunities"); *see also United States v. Fordice*, 505 U.S. 717, 731 (1992) ("If the State [university system] perpetuates policies and practices traceable to its prior system that continue to have segregative effects . . . and such policies are without sound educational justification and can be practically eliminated, the State has not satisfied its burden of proving that it has dismantled its prior system."); *Cf. GI Forum v. Texas Educ. Agency*, 87 F. Supp. 2d 667, 673, 684 (W.D. Tex. 2000) (the court concluded, based on the facts presented, that the test seeks to identify inequities and address them; the state had ensured that the exam is strongly correlated to material actually taught in the classroom; remedial efforts, on balance, are largely successful; and minority students have continued to narrow the passing gap).

B. Disparate Impact

The federal nondiscrimination regulations also provide that a recipient of federal funds may not "utilize criteria or methods of administration which have the effect of subjecting individuals to discrimination."[180] Thus, discrimination under federal law may occur where the application of neutral criteria is shown by the party challenging those criteria to have discriminatory effects and those criteria are not shown by the recipient to be educationally justified. Even if the criteria are educationally justified, discrimination may be found if it is shown by the challenging party that there are alternative practices available that are equally effective in serving the educational institution's goals and have less disparate impact. It is important to understand that disparities in student performance based on race, national origin, sex, or disability, do not alone constitute disparate impact discrimination under federal law; nothing in federal law guarantees equal results. Rather, significant disparities trigger further inquiry to ensure that the given policy is in fact nondiscriminatory.

Courts applying the disparate impact test have examined three questions to determine if the practice at issue is discriminatory: (1) Does the practice or procedure in question result in significant differences in the award of benefits or services based on race, national origin, or sex? (2) Is the practice or procedure educationally justified? and (3) Is there an equally effective alternative that can accomplish the institution's educational goal with less disparity?[181] (For a discussion of disability discrimination, including disparate impact discrimination, see discussion *infra* Chapter 2 (Legal Principles) Part III (Testing of Students with Disabilities).[182])

The party challenging the test has the burden of establishing disparate impact. If disparate impact is established, the educational institution must demonstrate the educational justification (also referred to as "educational necessity") of the practice in question.[183] If a sufficient educational justification is established, then the party challenging the test must establish that an alternative with less disparate impact is equally effective in meeting the institution's educational goals in order to prevail.[184]

---

[180] 34 C.F.R. § 100.3(b)(2) (Title VI); 34 C.F.R. § 104.4(b)(4)(i) (Section 504); 28 C.F.R. § 35.130(b)(3)(i) (Title II); *see also* 34 C.F.R. § 106.31 (Title IX). In *Guardians Association*, the U.S. Supreme Court upheld the use of the effects test, stating that the Title VI regulation forbids the use of federal funds, "not only in programs that intentionally discriminate on racial grounds but also in those endeavors that have a[n] [unjustified racially disproportionate] impact on racial minorities." 463 U.S. at 589-90.

[181] *Georgia State Conf. of Branches of NAACP v. Georgia*, 775 F.2d 1403, 1417 (11th Cir. 1985); *see also Elston*, 997 F.2d at 1407 n.14; *Larry P.*, 793 F.2d at 982 n.9; *Groves*, 776 F. Supp. at 1523-24, 1529-32; *Sharif*, 709 F. Supp. at 361. Many courts use the term "equally effective" when discussing whether the alternative offered by the party challenging the test is feasible and would effectively meet the institution's goals. *See, e.g., Georgia State Conf.*, 775 F.2d at 1417; *Sharif*, 709 F. Supp. at 361. Other courts use the term "comparably effective" in evaluating proposed alternatives. *See, e.g., Elston*, 997 F.2d at 1407; *Fitzpatrick v. City of Atlanta*, 2 F.3d 1112, 1118 (11th Cir. 1993). Review of the decisions in these cases indicates that the courts appear to be using the terms synonymously.

[182] Disparate impact disability discrimination may take forms that are not always amenable to analysis through the three-part approach usually applied in race or sex discrimination cases. For example, statistical evidence showing the effect of architectural barriers on persons of various types of disabilities may not be necessary. *See Choate*, 469 U.S. at 297-300. For this reason, disability discrimination is discussed separately. *See* discussion *infra* Chapter 2 (Legal Principles) Part III (Testing of Students with Disabilities).

[183] *Elston*, 997 F.2d at 1412.

1. Determining Disparate Impact

The first question in the disparate impact analysis is whether there is information indicating a significant disparity in the provision of benefits or services to students based on race, national origin, or sex. Courts have used a variety of methods to distinguish differences between outcomes that are statistically and practically significant from those that are random.[185] To determine if a sufficient disparate impact exists, courts have focused on evidence of statistical disparities.[186] Generally, a test has a disproportionate adverse impact if a statistical analysis shows a significant difference from the expected random distribution.[187]

> **Generally, if a statistical analysis shows that the success rate for a particular group of students is significantly lower (or the failure rate is significantly higher) than what would be expected from a random distribution, then the test has disproportionate adverse impact.**
>
> National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, p. 59 (Jay P. Heubert & Robert M. Hauser 1999).

There is no rigid mathematical threshold regarding the degree of disproportionality required; however, the statistical evidence must identify disparities that are sufficiently substantial to raise an inference that the challenged practice caused the disparate results.[188] To establish disparate impact in the context of a selection system, the comparison must be made between those selected for the educational benefit or service and a relevant pool of applicants or test takers.[189]

---

[184] *Georgia State Conf.*, 775 F.2d at 1417; *see also* Department of Justice, *Title VI Legal Manual*, p. 2.

[185] Different courts have used different methods for determining disparate impact. Some courts have used an 80 percent rule whereby disparate impact is shown when the rate of selection for the less successful group is less than 80 percent of the rate of selection for the most successful group. Another type of statistical analysis considers the difference between the expected and observed rates in terms of standard deviations, with the difference generally expected to be more than two or three standard deviations. Another test is known as the "Shoben formula" in which the difference or Z-value in the groups' success rates must be statistically significant. *Groves*, 776 F. Supp. at 1526-28 (discussing these methods and the cases in which they were used).

[188] *Watson*, 487 U.S. at 994-95; *Groves*, 776 F. Supp. at 1526-27.

[189] When determining disparate impact in the context of a selection system, the comparison pool generally consists of all minimally qualified test takers or applicants. When tests are used to determine placement or some other type of educational treatment, the comparison is between those identified by the test for the placement or educational treatment and the relevant pool of test takers. The precise composition of the comparison pool is determined on a case-by-case basis. *See Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 650-51 (1989); *Watson*, 487 U.S. at 995-97; *Groves*, 776 F. Supp. at 1525-26.

BEST COPY AVAILABLE

In general, a specific policy, practice, or procedure must be identified as causing the disproportionate adverse effect on the basis of race, national origin, or sex.[190] For example, when a particular use of a test is being challenged, the evidence should show that the test use, rather than other selection factors, accounts for the disparity.[191]

2. Determining Educational Necessity

Where the use of a test results in decisions that have a disparate impact on the basis of race, national origin, or sex, the test use causing the disparity must significantly serve the legitimate educational goals of the institution.[192] This inquiry is usually referred to as determining the "educational necessity" of the test use *or* determining whether the test is "educationally justified."[193]

In evaluating educational necessity, both the legitimacy of the educational goal asserted by the institution and the use of the test as a valid means to advance that goal may be at issue. Courts generally give deference to educational institutions to define their own legitimate educational goals[194] and focus more directly on whether the challenged test supports those goals.[195] While the test need not be "essential" or "indispensable" to achieving the institution's educational goal,[196] the educational institution must show a manifest relationship between use of the test and an important educational purpose.[197]

---

[190] As noted by Justice O'Connor in *Watson*, courts have found it "relatively easy," when appropriate statistical proof is presented, to identify a standardized test as causing the racial, national origin, or sex related disparity at issue. 487 U.S. at 994; see *also GI Forum*, 87 F. Supp. 2d at 677-79 (given legally meaningful differences in the pass rates of minority and majority students, plaintiffs made a prima facie showing of disparate impact resulting from a graduation test).

[191] Elements of a decision-making process that cannot be separated for purposes of analysis may be analyzed as one selection practice. *See* Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(k)(1)(B)(i). This is necessary because limiting the disparate impact analysis to a discrete component of a selection process would not allow for situations "where the adverse impact is caused by the interaction of two or more components of the process." *Graffam v. Scott Paper Co.*, 870 F. Supp. 389, 395 (D. Me. 1994), *aff'd*, 60 F.3d 809 (1995).

[192] *See Wards Cove*, 490 U.S. at 659.

[193] *See Board of Educ. v. Harris*, 444 U.S. 130, 151 (1979); *Elston*, 997 F.2d at 1412.

[194] *See Groves*, 776 F. Supp. at 1529 (citing *Wards Cove*, 490 U.S. at 659).

[195] *See, e.g., Debra P.*, 644 F.2d at 402 (indicating that the court is not in a position to determine education policy, and the state's efforts to establish minimum standards and improve educational quality are praiseworthy).

[196] *Wards Cove*, 490 U.S. at 659; *Elston*, 997 F.2d at 1412 (citing *Georgia State Conf.*, 775 F.2d at 1417-18).

[197] See *Georgia State Conf.*, 775 F.2d at 1418 (showing required that "achievement grouping practices bear a manifest demonstrable relationship to classroom education"); *Sharif*, 709 F. Supp. at 362 (defendants must show a manifest relationship between use of the SAT and recognition of academic achievement in high school). As explained in *Elston*, "from consulting the way in which . . . [courts] analyze the 'educational necessity' issue, it becomes clear that . . . [they] are essentially requiring . . . [the educational institution to] show that the challenged course of action is demonstrably necessary to meeting an important educational goal." *Elston*, 997 F.2d at 1412. In other words, the institution can defend the challenged practice on the grounds that it is "supported by a 'substantial legitimate justification.'" *Id.* (quoting *Georgia State Conf.*, 775 F.2d at 1417); *see, e.g., Georgia State Conf.*, 775 F.2d at 1417-18; *Groves*, 776 F. Supp. at 1529-32.

In conducting this analysis, courts have generally considered relevant evidence of validity, reliability, and fairness[198] provided by the test developer and test user to determine the acceptability of the test for the purpose used, giving deference, as appropriate, to the educational institution's testing practices that are within professionally accepted standards.[199] The educational justification inquiry thus generally looks at technical questions regarding the test's accuracy in relation to the nature and importance of the educational institution's goals, the educational consequences to students, the relationship of the educational institution to the student, and other factors bearing on test use, such as whether and how additional information beyond the test score enters into the educational decision at stake.[200]

---

[198] In general, courts have said that validity refers to the accuracy of conclusions drawn from test results. *See Allen v. Alabama State Bd. of Educ.*, 976 F. Supp. 1410, 1420-21 (M.D. Ala. 1997) ("Generally, validity is defined as the degree to which a certain inference from a test is appropriate and meaningful," quoting *Richardson v. Lamar County Bd. of Educ.*, 729 F. Supp. 809, 820 (M.D. Ala. 1989), *aff'd*, 164 F.3d 1347 (11th Cir. 1999), *injunction granted*, 2000 U.S. Dist. LEXIS 123 (M.D. Ala.)); *see also Richardson*, 729 F. Supp. at 820-21 ("[A] test will be valid so long as it is built to yield its intended inference and the design and execution of the test are within the bounds of professional standards accepted by the testing industry."); *Anderson v. Banks*, 520 F. Supp. 472, 489 (S.D. Ga. 1981) ("Validity in the testing field indicates whether a test measures what it is supposed to measure.").

[199] *See, e.g., United States v. LULAC*, 793 F.2d 636, 640, 649 (5th Cir. 1986) (pointing to substantial expert evidence in the record, including validity studies, indicating that the tests involved were valid measures of the basic skills that teachers should have). The sponsors of the newly revised *Joint Standards* advise that the *Joint Standards* is intended to provide guidance to testing professionals in making such judgments. *Joint Standards, supra* note 3, at p.4. The *Joint Standards* is discussed more fully in Chapter One of this guide.

Where the evidence indicates that the educational institution is using a test in a manner that does not lead to valid inferences, educational justification may be found lacking. *Groves*, 776 F. Supp. at 1530 (requiring minimum ACT score for admission to undergraduate teacher education programs violated the Title VI regulations since ACT scores had not been validated for this purpose); *Sharif*, 709 F. Supp. at 361-63 (in ruling on a motion for preliminary injunction, court found that the state's use of SAT scores as the sole basis for decisions awarding college scholarships intended to reward high school achievement was not educationally justified for this purpose in that the SAT had been designed as an aptitude test to predict college success and was not designed or validated to measure past high school achievement); *See Fordice*, 505 U.S. at 736-37 (ruling that Mississippi's exclusive use of ACT scores in making college admissions decisions was not educationally justified, since, among other factors, the ACT's administering organization discouraged this practice).

Numeric evidence is not the only way that validity can be demonstrated, however. Courts can draw inferences of validity from a wide range of data points. *Watson*, 487 U.S. at 998 (referring to procedures used to evaluate personal qualities of candidates for managerial jobs).

[200] *See, e.g., Larry P.*, 793 F.2d at 980; *Georgia State Conf.*, 775 F.2d at 1417-20; *Groves*, 776 F. Supp. at 1530-31. In the educational context, tests play a complex role that bears on evaluation of educational justification. As noted by the court in *Larry P.*,

> [I]f tests can predict that a person is going to be a poor employee, the employer can legitimately deny that person a job, but if tests suggest that a young child is probably going to be a poor student, the school cannot on that basis alone deny that child the opportunity to improve and develop the academic skills necessary to success in our society.

793 F.2d at 980 (quoting *Larry P., v. Riles*, 495 F. Supp. 926, 969 (1979)). Because determining whether a test is a valid basis for classifying students and placing them in different educational programs may be even more complex and difficult than determining if a test validly predicts job performance, particular sensitivity is needed to all of the interests involved. The question may be not only whether a test provides valid information about a student's ability and achievement, but whether the educational services provided to the student as a consequence of the test serve the student's needs. Inequality in the services provided to students prior to the test, as well as in the services provided as a consequence of the test, may also be a factor considered as part of the educational justification for

Where a test is used for promotion or graduation purposes, a major consideration is the extent to which the educational institution has provided the student with the opportunity to learn the content and skills being tested.[201]

3. Determining Whether There Are Equally Effective Alternatives that Serve the Institution's Educational Goal with Less Disparity

If the educational institution provides sufficient evidence that the test use in question is justified educationally, the party challenging the test has the opportunity to show that there exists an equally effective alternative practice that meets the institution's goals with less disparity.[202] The feasibility of an alternative, including costs and administrative burdens, is a relevant consideration.[203]

II. Testing Of Students With Limited English Proficiency

Testing of students with limited English proficiency in the elementary and secondary education context raises a set of unique issues. To understand the obligations of states and school districts with regard to high-stakes testing of such students, it is important to understand the basic obligations of school districts and states under Title VI and federal law that relate to language minority students who are learning English.

Title VI prohibits discrimination based on race, color, or national origin. On May 25, 1970, the United States Department of Health, Education, and Welfare's Office for Civil Rights issued a policy memorandum entitled "Identification of Discrimination and Denial of Services on the Basis of National Origin." The May 25th memorandum clarified the

---

using a test in a particular way. *See Debra P.*, 644 F.2d at 407-08 (agreeing with the statement that Title VI would not be violated if the test were a fair test of what students were taught); *Debra P. v. Turlington*, 730 F.2d 1405, 1407, 1410-11, 1416 (11th Cir. 1984) (affirming that the extent of remedial efforts to address test failure is relevant to evaluation of test use).

[201] *See Debra P.*, 644 F.2d at 408.

[202] *New York Urban League v. New York*, 71 F.3d 1031, 1036 (2d Cir. 1995) (stating "the plaintiff may still prove his case by demonstrating that other less discriminatory means would serve the same objective"); *see also Albemarle Paper Co. v. Moody*, 422 U.S. 405, 425 (1975); *Richardson*, 729 F. Supp. at 815. Alternative practices that have been offered for examination include procedures that consider additional types of performance information along with test results consistent with the institution's goals. *See, e.g.*, *Sharif*, 709 F. Supp. at 362063 (consideration of SAT score plus grade point average would be a better measure of high school achievement for purpose of scholarship eligibility than SAT score alone); *GI Forum*, 87 F. Supp. 2d at 681 (consideration of grades along with graduation test scores would not further state's legitimate purpose in using test).

[203] *See Wards Cove*, 490 U.S. at 661 (indicating that factors such as costs or other burdens are relevant in determining whether the alternative is equally effective in serving employer's legitimate goals); *MacPherson v. University of Montevallo*, 922 F.2d 766, 773 (11th Cir. 1991) (holding that plaintiff must show that the alternative is economically feasible); *Sharif*, 709 F. Supp. at 363-64 (finding defendant's claim that proposed alternative was not feasible and excessively burdensome not persuasive since most other states used proposed alternative).

responsibility of school districts, under Title VI, to provide equal educational opportunity to national origin minority group students whose inability to speak and understand the English language excludes them from effective participation in any education program offered by a school district.[204] This memorandum was cited with approval by the Supreme Court in its decision in *Lau v. Nichols*, which held that the district's policy of teaching national origin minority group children only in English, without any special assistance, deprived them of the opportunity to benefit from the district's education program, including meeting the English language proficiency standards required by the state for a high school diploma.[205] The *Lau* case held that such policies are barred when they have the effect of denying such benefits, even though no purposeful design is present.[206]

Subsequently, *Castaneda v. Pickard*,[207] relying on the language of the Equal Educational Opportunities Act (EEOA), explained the steps school districts must take to help students with limited English proficiency overcome language barriers to ensure that they can participate meaningfully in the districts' educational programs.[208] The court stated that school districts have an obligation to provide services that enable students to acquire English language proficiency. A school system that chooses to temporarily emphasize English over other subjects retains an obligation to provide assistance necessary to remedy academic deficits that may have occurred in other subjects while the student was focusing on learning English.

Under the *Castaneda* standards, school districts have broad discretion in choosing a program of instruction for limited English proficient students. However, the program must be based on sound educational theory, must be adequately supported so that the program has a realistic chance of success, and must be periodically evaluated and revised, if necessary, to achieve its goals.

The disparate impact framework discussed earlier in the guide in Chapter 2 Part (I)(B) may also be used to examine whether tests used for high-stakes purposes result in a discriminatory impact upon students with limited English proficiency. As part of this analysis, questions may arise regarding the validity and reliability of the test for these students.[209]

---

[204] Identification of Discrimination and Denial of Services on the Basis of National Origin, 35 Fed. Reg. 11595 (1970). The Department of Health, Education and Welfare was the predecessor of the U.S. Department of Education.

[205] *Lau*, 414 U.S. at 566-68.

[206] *Lau*, 414 U.S. at 568 (citing, among other legal authority, the predecessor of 34 C.F.R. § 100.3 (b)(2)).

[207] *Castanada v. Pickard*, 648 F.2d 989, 1005-06, 1009-12 (5th Cir. 1981). The analytical framework in *Castaneda* which was decided under the Equal Educational Opportunities Act (EEOA), 20 U.S.C. §§ 1701 *et seq.*, has been applied to OCR's Title VI analysis. *See Williams Memorandum, supra* note 50. The EEOA contains standards related to limited English proficient students similar to the Title VI regulations.

[208] *Castaneda*, 648 F.2d at 1011.

[209] *See* discussion *supra* Chapter 1 (Test Measurement Principles) Part (II)(B) (Testing of Limited English Proficient Students) for a discussion of the relevant principles involved in determining the reliability and validity of tests used with limited English proficient students.

Depending upon the purpose of the test and the characteristics of the populations being tested, in some situations, accommodations or other forms of assessment of the same construct may be necessary. In short, the obligation is to ensure that the same constructs are being measured for all students.

There are three particularly important areas involving high-stakes testing of students with limited English proficiency: (1) tests used to determine a student's proficiency in the areas of speaking, listening, reading, or writing English for the purpose of determining whether the student should be provided with a program or services to enable the student to acquire English language skills (and, later, for the purpose of determining whether the student is ready to exit the program or services); (2) tests used to determine if the student meets the criteria for other specialized instructional programs, such as gifted and talented or vocational education programs; and (3) systemwide tests, including graduation tests, administered to determine if students have met performance standards.

Tests used to determine a student's initial and continuing need for special language programs should be appropriate in light of a district's own performance expectations and otherwise valid and reliable for the purpose used. Tests used by schools to help select students for specialized instructional programs, including programs for gifted and talented students, should not screen out limited English proficient students unless the program itself requires proficiency in English for meaningful participation.[210] When a state or school district adopts content and performance standards and uses tests for high-stakes purposes, such as graduation tests, to measure whether students have mastered those standards, a critical factor under Title VI is whether the overall educational program provided to students with limited English proficiency is reasonably calculated to enable the students to master the knowledge and skills that are required to pass the test. When education agencies institute standards-based testing, it is important for them to examine their programs for students with limited English proficiency to determine when and how these students will be provided with the instruction needed to prepare them to pass the test in question.[211]

In addition, students with limited English proficiency may not be categorically excluded from standardized testing designed to increase accountability of educational programs for effective instruction and student performance. If these students are not included, the test data will not fairly reflect the performance of all students for whom the education agency is responsible.[212] Such test data can also help a district assess the effectiveness of its content and English language acquisition programs.

---

[210] *Williams Memorandum, supra*, note 50.

[211] Careful attention to the alignment between instructional content and testing standards is especially important for students who receive instruction that deviates from the regular curriculum. *See Brookhart v. Illinois State Bd. of Educ.*, 697 F.2d 179, 186-87 (7th Cir. 1982) (finding that students with disabilities in special education programs were denied exposure to most of the material covered in a newly instituted graduation test).

[212] Indeed, Title I of the Elementary and Secondary Education Act explicitly requires states to include limited English proficient students in the statewide assessments used to hold schools and school districts accountable for student performance. Title I of the Elementary and Secondary Education Act, 20 U.S.C. § 6311(b)(3)(F)(iii). If a school district uses the results of a test given for program accountability purposes to make educational decisions about individual students, the high-stakes use of the test must also be valid and reliable for this purpose.

For information on the factors that help ensure accuracy of tests for limited English proficient students, see discussion *infra* Chapter 1 (Test Measurement Principles) Part II (B) (Testing of Limited English Proficient Students). In making decisions about testing limited English proficient students, factors such as the student's level of English proficiency, the primary language of instruction, the level of literacy in the native language, and the number of years of instruction in English may all be pertinent.[213] When students participate in assessments designed to meet the requirements of Title I of the Elementary and Secondary Education Act, as amended, those assessments must be implemented in a manner that is consistent with both the requirements of Title VI and Title I.

III. Testing Of Students With Disabilities

Three federal statutes provide basic protections for elementary and secondary students with disabilities. Section 504 of the Rehabilitation Act of 1973 (Section 504) and Title II of the Americans with Disabilities Act of 1990 (Title II) prohibit discrimination against persons with disabilities by public schools.[214] The Individuals with Disabilities Education Act (IDEA) establishes rights and protections for students with disabilities and their families. It also provides federal funds to state education agencies and school districts to assist in educating students with disabilities.[215] Under Section 504, Title II, and the IDEA,[216] school districts have a responsibility to provide students with disabilities, as defined by applicable law, with a free appropriate public education. Providing effective instruction in the general curriculum for students with disabilities is an important aspect of providing a free appropriate public education.

The regulations implementing Section 504 and Title II specifically prohibit the use of "criteria or methods of administration . . . that have the effect of subjecting qualified persons with disabilities to discrimination on the basis of disability."[217] Under Section 504, Title II, and the IDEA, tests given to students with disabilities must be selected and administered so that the test accurately reflects what a student knows or is able to do, rather than a student's disability (except when the test is designed to measure disability-related skills). This means that students with disabilities covered by these statutes must be given appropriate accommodations and modifications in the administration of the tests that allow the same constructs to be measured. Examples include oral testing, tests in large print, Braille versions of tests, individual testing, and separate group testing.

---

[213] For more information on appropriate practices for testing students who are learning English, *see* Kopriva, *Ensuring Accuracy in Testing, supra* note 131.

[214] Although this part of the chapter deals only with students with disabilities attending public elementary and secondary schools, private schools that are not religious schools operated by religious organizations are covered by Title III of the ADA. Title III of the Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12181 *et seq*. In addition, Title I of the Elementary and Secondary Education Act of 1965, as amended, contains important provisions regarding students with disabilities in the Title I program and their participation in assessments of Title I programs. *See* 20 U.S.C. § 6311(b)(3)(F).

[215] Individuals with Disabilities Education Act, 20 U.S.C. § 1400(d)(1)(c).

[216] The Section 504 regulation is found at 34 C.F.R. Part 104. The Title II regulation is found at 28 C.F.R. Part 35. The IDEA regulation is found at 34 C.F.R. Part 300.

[217] 28 C.F.R. § 35.130(b)(3); 34 C.F.R. § 104.4(b)(4). In *Guardians Association*, the United States Supreme Court upheld the use of the effects test in the context of Title VI, stating that the Title VI regulation forbids the use of federal funds, "not only in programs that intentionally discriminate on racial grounds but also in those endeavors that have a [racially disproportionate] impact on racial minorities." 463 U.S. at 589.

Generally, there are three critical areas in which high-stakes testing issues arise for students with disabilities: (1) tests used to determine whether a student has a disability and, if so, the nature of the disability; (2) tests used to determine if a student meets the criteria for other specialized instructional programs, such as gifted and talented or vocational education programs; and (3) systemwide tests administered to determine if a student has met performance standards.

Under Section 504, Title II, and the IDEA, before an elementary and secondary school student can be classified as having a disability, the responsible education agency must individually evaluate the student in accordance with specific statutory and regulatory requirements, including requirements regarding the validity of tests and the provision of appropriate accommodations.[218] These requirements prohibit the use of a single test score as the sole criterion for determining whether a student has a disability and for determining an appropriate educational placement for the student.[219]

When tests are used for other purposes, such as in making decisions about placement in gifted and talented programs, it is important that tests measure the skills and abilities needed in the program, rather than the disability, unless the test purports to measure skills or functions that are impaired by the disability and such functions are necessary for participation in the program.[220] For this reason, appropriate accommodations may need to be provided to students with disabilities in order to measure accurately their performance in the skills and abilities required in the program.

Furthermore, federal laws generally require the inclusion of students with disabilities in state- and districtwide assessment programs, except as participation in particular tests is individually determined to be inappropriate for a particular student. Assessment programs should provide valuable information that benefits students, either directly, such as in the measurement of individual progress against standards, or indirectly, such as in evaluating programs. Given these benefits, exclusion from assessment programs, unless such participation is individually determined inappropriate because of the student's disability, would generally violate Section 504 and Title II. If a student with a disability will take the systemwide assessment test the student must be provided appropriate instruction and appropriate test accommodations.[221] The Individuals with Disabilities Education Amendments of 1997 specifically require states, as a condition of receiving IDEA funds, to include students with disabilities in the regular state- and districtwide assessment programs, with appropriate accommodations, where necessary.[222] The IDEA also requires state or local education agencies to develop guidelines for the relatively small number of students with disabilities who cannot take part in state- and districtwide tests to participate in alternate assessments.[223]

---

[218] *See* 34 C.F.R. § 104.35(b) (specific provisions covering the use of tests for evaluation purposes).

[219] *See* 34 C.F.R. § 104.35(c) (requiring placement decisions to consider information from a variety of sources).

[220] 34 C.F.R. §§ 104.35(b)(3), 300.532.

For children with disabilities, school personnel knowledgeable about the student, the nature of the disability, and the testing program, in conjunction with the student's parent or guardian, determine whether the student will participate in all or part of the state- or districtwide assessment of student achievement.[224] The decision must be documented in the student's individualized education program (IEP), or a similar record, such as a Section 504 plan. These records must also state any individual accommodations in the administration of the state- or districtwide assessments of student achievement that are needed to enable the student to participate in such assessment. An IEP, developed under the IDEA, must also explain how the student will be assessed if it is inappropriate for the student to participate in the testing program even with accommodations.[225] The individual decisions made regarding testing of the student in the IEP or Section 504 plan are subject to appeal by the parent or guardian through the due process procedures required by applicable law.[226]

Section 504 and Title II also prohibit discrimination against qualified persons with disabilities in virtually all public and private post-secondary institutions.[227] The regulatory requirements related to disability discrimination are different in post-secondary education than in elementary and secondary education. Post-secondary institutions are not required to evaluate students or to provide them with a free appropriate education.

---

[221] *Brookhart*, 697 F.2d at 183-84. Some courts have held that a student with a disability may be denied a diploma if, despite receiving appropriate services and testing accommodations, the student, because of the disability, is unable to pass the required test or meet other graduation requirements. *Id*., at 183; *Anderson*, 520 F. Supp. at 509-11; *Board of Educ. v. Ambach*, 458 N.Y.S.2d 680, 684-85, 689 (N.Y. App. Div. 1982), *aff'd*, 469 N.Y.S.2d 669 (1983).

[222] 20 U.S.C. § 1412(a)(17); 34 C.F.R. § 300.138(a).

[223] 34 C.F.R. § 300.138(b). The IDEA Final Regulations, Attachment I—Analysis of Comments and Changes, 64 Fed. Reg. 12406, 12564 (1999), projects that there will be a relatively small number of students who will not be able to participate in the district or state assessment program with accommodations and modifications, and will therefore need to be assessed through alternate means. These alternate assessments must be developed and conducted beginning not later than July 1, 2000.

[224] *See* 34 C.F.R. § 300.347(a)(5) (IEP requirements applicable to assessment of students with disabilities under IDEA); 34 C.F.R. § 104.33 (more general evaluation requirements under Section 504).

[225] 34 C.F.R. § 300.347(a)(5).

[226] 34 C.F.R. §§ 300.507, 104.36.

[227] Under the Section 504 regulation, a qualified person with a disability for purposes of post-secondary education is an individual with a disability within the meaning of the regulation who meets the academic and technical standards for admission. 34 C.F.R. §§ 104.3(j), 104.3(k).

High-stakes testing issues at the post-secondary level generally relate to tests considered by post-secondary institutions for admissions, including tests given by an educational institution or other covered entities as prerequisites for entering a career or career path, and tests of academic competency required by the institution to complete a program. This guide is not intended to offer a complete or detailed explanation of each of these testing situations, but only a brief synopsis.[228]

The Section 504 regulation specifically provides that higher education institutions' admissions procedures may not make use of any test or criterion for admission that has a disproportionate, adverse impact on individuals with disabilities unless (1) the test or criterion, as used by the institution, has been validated as a predictor of success in the education program or activity and (2) alternative tests or criteria that have a less disproportionate, adverse impact are not shown to be available.[229] In administering tests, appropriate accommodations must be provided so that the person can demonstrate his or her aptitude and achievement, not the effect of the disability (except where the functions impaired by the disability are the factors the test purports to measure).[230]

For other high-stakes tests that an institution might administer, such as rising junior tests, similar requirements apply.[231] The institution must provide adjustments or accommodations and auxiliary aids and services that enable the student to demonstrate the knowledge and skills being tested.[232]

Students are required to notify the educational institution when accommodations are needed and initially supply adequate documentation of a current disability and the need for accommodation.[233] The student's preferred accommodation does not have to be provided as long as an effective accommodation is provided.

Test accommodations are intended to provide the person with disabilities the means by which to demonstrate the skills and knowledge being tested. Although Section 504 and Title II require a college or university to make reasonable modifications, neither Section 504 nor Title II requires a college or university to change, lower, waive, or eliminate academic requirements or technical standards that can be demonstrated by the college or university to be essential to its program of instruction or to any directly related licensing requirement.[234] Accommodations requested by students need not be provided if they would result in a fundamental alteration to the institution's program.[235]

---

[228] Test providers that are not higher education institutions may be covered by Section 504 if they receive federal funds; by Title II if they are parts of governmental units; or by Title III if they are private entities. Each of these laws has its own requirements. For more information regarding testing under Title III of the ADA, consult the U.S. Department of Justice.

[229] 34 C.F.R. § 104.42(b)(2). Appendix A to the Section 504 regulation, Subpart E-Post-secondary Education, No. 29, notes that the party challenging the test would have the burden of showing that alternate tests with less disparate impact are available.

[230] 34 C.F.R. § 104.42(b)(3).

[231] Some undergraduate college programs require students to pass a rising junior examination to determine whether students have met the college's standards in writing or other academic skills as a prerequisite for advancement to junior year status.

[232] 34 C.F.R. §§ 104.44(a), 104.44(d).

[233] See, *e.g.*, *Kaltenberger v. Ohio College of Podiatric Medicine*, 162 F.3d 432, 437 (6th Cir. 1998).

IV. Constitutional Protections

In addition to applying federal nondiscrimination statutes, courts have also considered constitutional issues that may arise when public school districts or state education agencies utilize tests for high-stakes purposes in their educational programs, particularly tests required for promotion or graduation.[236] Constitutional challenges to testing programs under the Fourteenth Amendment have raised both equal protection and due process claims. The equal protection principles involved in discrimination cases are, generally speaking, the same as the standards applied to intentional discrimination claims under the applicable federal nondiscrimination statutes.[237]

---

[234] *See* 34 C.F.R. § 104.44(a).

[235] *Southeastern Community College v. Davis*, 442 U.S. 397, 413 (1979); *Wynne v. Tufts Univ. Sch. of Medicine*, 976 F.2d 791, 794-96 (1st Cir. 1992), *cert. denied*, 507 U.S. 1030 (1993).

[236] The U.S. Department of Education, Office for Civil Rights, does not have jurisdiction to resolve constitutional cases. However, some cases involve constitutional issues that overlap with discrimination issues arising under federal civil rights laws.

[237] Federal cases may also involve equal protection challenges to a jurisdiction's use of tests in which the claim is not based on race or sex discrimination, but instead on the alleged impropriety of the jurisdiction's use of the test in making educational decisions. As a general matter, courts express reluctance to second guess a state's educational policy choices when faced with such challenges, although recognize that a state cannot "exercise that [plenary] power without reason and without regard to the United States Constitution." *Debra P.*, 644 F.2d at 403. When there is no claim of discrimination based on membership in a suspect class, the equal protection claim is reviewed under the rational basis standard. In these cases, the jurisdiction need show only that the use of the tests has a rational relationship to a valid state interest. *Id.*, at 406; *see also Erik V. v. Causby*, 977 F. Supp. 384, 389 (E.D.N.C. 1997).

The due process clause of the Fourteenth Amendment is particularly associated with cases challenging the adequacy of the notice provided to students prior to this type of test and the students' opportunity to learn the required content.[238] In analyzing such due process claims, courts have generally considered three issues:

(1) Is the testing program reasonably related to a legitimate educational purpose?

Federal courts typically defer to educators' policy judgments regarding the value of legitimate educational benefits sought from the testing programs.[239] For example, improving the quality of elementary and secondary education through the establishment of academic standards has been seen as a legitimate goal of a testing program, and colleges and universities generally have been given wide latitude in framing degree requirements.[240] The constitutional inquiry then focuses on whether the challenged testing program is

---

[238] A review of relevant cases reveals the highly fact- and context-specific nature of the conclusions reached by federal courts considering alleged violations of the due process clause. In *Debra P.*, the Fifth Circuit held that students' due process rights were violated when a newly imposed minimum competency test required for high school graduation was instituted without adequate notice and an opportunity for students to learn the material covered by the test. 644 F.2d at 404. Three years later, in *Debra P. v. Turlington*, the court held that students who now had six years notice of the exam were afforded the opportunity to learn the relevant material, given the state's remedial programs. 730 F.2d at 1416-17. For additional courts identifying due process violations in the way in which a competency test was instituted, *see Brookhart*, 697 F.2d at 186-87 (holding that district-required minimum competency test for graduation denied due process to students with disabilities where notice was inadequate and students had not been exposed to 90 percent of the material covered by the test); *Crump v. Gilmer Indep. Sch. Dist.*, 797 F. Supp. 552, 556-57 (E.D. Tex. 1992) (granting temporary restraining order where district had not demonstrated validity of graduation examination in light of actual instructional content); *Anderson*, 520 F. Supp. at 508-09 (finding that school district failed to show that minimum competency test required for high school graduation covered material actually taught at school). Other cases have concluded that adequate notice was provided, the test or criterion at issue was closely related to the instructional program, or the promotion decision was not shown to be outside the discretion of school authorities. *See Erik V.*, 977 F. Supp. at 389-90 (finding that promotion decision was within proper purview of school authorities); *Williams v. Austin Indep. Sch. Dist.*, 796 F. Supp. 251, 253-54 (W.D. Tex. 1992) (considering students to have had seven years advance notice of high school competency exam although standards of performance were recently raised). Also relevant are promotion cases in which students were required to demonstrate adequate reading skills, although a separate test was not apparently involved. *See Bester v. Tuscaloosa City Bd. of Educ.*, 722 F.2d 1514, 1516 (11th Cir. 1984) (finding reading standards required for promotion to merely reinforce district policy of retention for substandard work); *Sandlin v. Johnson*, 643 F.2d 1027, 1029 (4th Cir. 1981) (finding denial of second-grade promotion for failing to attain required level in reading series within discretion of school district). For a testing case raising similar due process issues at the post-secondary level, *see Mahavongsanan v. Hall*, 529 F.2d 448, 450 (5th Cir. 1976) (finding no violation of due process where the university's decision to require a comprehensive examination for receipt of a graduate degree was a reasonable academic regulation, plaintiff received timely notice that she would be required to take the examination, she was allowed to retake the test, and the university afforded her an opportunity to complete additional course work in lieu of the examination).

[239] See *Regents of the Univ. of Mich. v. Ewing*, 474 U.S. 214, 226-27 (1985); *Debra P.*, 644 F.2d at 406; *Anderson*, 520 F. Supp. at 506.

[240] *See Ewing*, 474 U.S. at 222, 226-27. (acknowledging that courts will not review academic decisions of colleges and universities unless the decision is such a substantial departure from accepted academic norms as to demonstrate that professional judgment was not actually exercised or where discrimination is claimed); *Debra P.*, 644 F.2d at 402 (finding praiseworthy a state's effort to set standards to improve public education).

BEST COPY AVAILABLE

reasonably related to the educators' legitimate goals or whether the program produces results that are arbitrary and capricious or fundamentally unfair.[241]

(2) Have students received adequate notice of the test and its consequences?

In the elementary and secondary school context, courts have required sufficient advance notice of tests required for graduation to give students a reasonable chance to learn the material presented on the test.[242] A particularly important concern in some of these decisions is the adequacy of notice provided to students. This issue has arisen in cases where racial minority students and students with disabilities received inadequate notice and did not receive a program of instruction that prepared them to pass the test.[243] In looking at the length of the transition period needed between the announcement of a new requirement and its full implementation, the kind of test and the context in which it is administered are central factors to be considered. Specific circumstances taken into account include the nature of instructional supports, including remediation, that accompany the test,[244] whether re-testing is permitted,[245] and whether the decision to promote or graduate the student considers other information about the student's performance.[246]

---

[241] The determination as to whether a testing program is rationally related to a legitimate educational goal has been considered under the Fourteenth Amendment as an issue of substantive due process. *See Debra P.*, 644 F.2d at 404-06; *Anderson*, 520 F. Supp. at 506. Insofar as due process cases may involve other technical questions of the validity of the test used to address the institution's goals, these issues are discussed in the portions of the guide addressing discrimination under federal civil rights laws.

[242] Although there are important exceptions, (*United States v. LULAC*, 793 F.2d 636, 648 (5th Cir. 1986), and *Anderson*, 520 F. Supp. at 505), courts have often considered the issue of adequate notice to be one of procedural due process. For procedural due process to apply, a protected property or liberty interest must be identified. *See Brookhart*, 697 F.2d at 185 (identifying a liberty interest, based on stigma of diploma denial, that disastrously affected plaintiffs' future employment and educational opportunities); *Debra P.*, 644 F.2d at 404 (finding sufficient to trigger due process protection a state-created mutual expectation that students who successfully complete required courses would receive diploma); *Erik V.*, 977 F. Supp. at 389-90 (finding no property interest in grade-level promotion warranting preliminary injunction).

[243] *See Brookhart*, 697 F.2d at 186-88; *Debra P.*, 644 F.2d at 404.

[244] *See Debra P.*, 730 F.2d at 1407, 1410-12, 1415-16; *Anderson*, 520 F. Supp. at 505.

[245] Re-testing was available in *Erik V.*, 977 F. Supp. at 388-89, and in *Anderson*, 520 F. Supp. at 505.

[246] *See Erik V.*, 977 F. Supp. at 387 (reading performance of students with grades of A, B, or C on grade-level work was further reviewed by teacher and principal to determine if student should be promoted notwithstanding the failing test score).

(3) Are students actually taught the knowledge and skills measured by the test?

Several courts have found that "fundamental fairness" requires that students be taught the material covered by the test where passing the test is a condition for receipt of a high school diploma.[247] For example, in analyzing this issue in a case involving a state where there had been past intentional segregation in elementary and secondary schools before a statewide diploma test was required, and where racial minority students had a disproportionate failure rate on the test, the court took the state's past intentional segregation into account in determining whether racial minority students had been given opportunities to learn the material covered by the test.[248] For the test to meaningfully measure student achievement, the test, the curriculum, and classroom instruction should be aligned. In cases examining systemwide administration of a test, courts require evidence that the content covered by the test is actually taught, but may not expect proof that every student has received the relevant instruction.[249]

---

[247] The question of opportunity to learn (sometimes called instructional or curricular validity) may be posed as one of substantive due process. *See Debra P.*, 644 F.2d at 406; *Anderson*, 520 F. Supp. at 509.

[248] *See Debra P.*, 644 F.2d at 407 (where black students disproportionately failed a statewide test necessary to obtain a high school diploma, and, due to the prior dual school system, black students received a portion of their education in unequal, inferior segregated schools, and where the state was unable to show that the diploma sanction did not perpetuate the effects of that past intentional discrimination, the court found that immediate use of the diploma sanction punished the black students for deficiencies created by the dual school system in violation of their constitutional right to equal protection); *Debra P.*, 474 F. Supp. 244, 257 (M.D. Fla. 1979) ("punishing the victims of past discrimination for deficits created by an inferior educational environment neither constitutes a remedy nor created better educational opportunities").

[249] *See Anderson v. Banks*, 540 F. Supp. 761, 765 (S.D. Ga. 1982).

APPENDIX A: Glossary of Legal Terms

This glossary is provided as a plain language reference to assist non-lawyers in understanding commonly used legal terms that are either used in this guide or are important to know in understanding the terms in the guide. Legal terms are often "terms of art." In other words, they mean something slightly different or more specific in the legal context than they do in ordinary conversation.

Burden of proof—the duty of a party to substantiate its claim or defense against the other party. In civil actions, the weight of this proof is usually described as a preponderance of the evidence. BLACK'S LAW DICTIONARY 196-97 (6th ed. 1990); *see also Disparate impact.*

Constitutional rights—the rights of each American citizen that are guaranteed by the United States Constitution. *See Brown v. Board of Educ., 347 U.S. 483 (1954); Bolling v. Sharpe*, 347 U.S. 497 (1954); BLACK'S LAW DICTIONARY 312 (6th ed. 1990).

***De jure* segregation or discrimination**— term applied to systemic school segregation that was mandated by statute or that was accomplished through the intentionally segregative actions of local school boards or state education agencies.

Different treatment—a claim that similarly situated persons are treated differently because of their race, color, national origin, sex or disability. Under federal nondiscrimination laws, policies and practices must be applied consistently to an individual or group of students regardless of their race, national origin, sex, or disability, unless there is a legally permissible reason for not doing so. Title VI, Title IX, Section 504, and the ADA prohibit intentional discrimination on the basis of race, national origin, color, sex, or disability. *Elston v. Talladega County Bd. of Educ.*, 997 F.2d 1394, 1406 (11th Cir. 1993). This requires a showing that the decision-maker was not only aware of the person's race, national origin, sex, or disability, but that the recipient acted, at least in part, because of the person's race, national origin, sex, or disability. However, the record need not contain "direct evidence of bad faith, ill will or any evil motive," on the part of the recipient. *Id.*, at 1406 (quoting *Williams v. City of Dotham*, 745 F.2d 1406, 1414 (11th Cir. 1984)). Evidence of discriminatory intent may be direct or circumstantial such as evidence of different treatment. Different treatment may be justified by a lawful reason, for example, to remedy prior discrimination. *See generally United States v. Fordice*, 505 U.S. 717, 728-30 (1992); *Wygant v. Jackson Bd. of Educ.*, 476 U.S. 267, 290-91 (1986); *Regents of the Univ. of Cal. v. Bakke*, 438 U.S. 265, 305-20 (1978); *Hopwood v. Texas*, 78 F.3d 932, 948-50 (5th Cir. 1996), *cert. denied*, 518 U.S. 1033 (1996); BLACK'S LAW DICTIONARY 470 (6th ed. 1990).

Disparate impact—disparate impact analysis applies when the application of a neutral criterion or a facially neutral practice has discriminatory effects and the criterion or practice is not determined to be "educationally justified" or "educationally necessary." In contrast to intentional discrimination, the disparate impact analysis does not require proof of discriminatory motive. Under the disparate impact analysis, the party challenging the criterion or practice has the burden of establishing disparate impact. If disparate impact is established, the party defending the practice must establish an "educational justification." If the educational institution provides sufficient evidence that the test use in question is justified educationally, the party challenging the test has the opportunity to show that there exists an alternative practice that meets the institution's goals as well as the challenged test use and that would eliminate or reduce the adverse impact. *See Board of Educ. v. Harris*, 444 U.S. 130, 143 (1979); *Georgia State Conf. of Branches of NAACP v. Georgia*, 775 F.2d 1403, 1412 (11th Cir. 1985); *Groves v. Alabama State Bd. of Educ.*, 776 F. Supp. 1518 (M.D. Ala. 1991).

Dual system—a previously segregated educational system in which black and white schools, ostensibly similar, existed side-by-side. *See Brown v. Board of Educ.*, 347 U.S. 483 (1954); *Anderson v. Banks*, 520 F. Supp. 472, 499-501 (S.D. Ga. 1981).

Due process—a constitutionally guaranteed right. The Fifth Amendment states that no citizen shall "be deprived of life, liberty, or property, without due process of law." The Fourteenth Amendment applied this passage to the states as well. Today it is used by the judiciary to define the scope of fundamental fairness due to each citizen in his or her interactions with the government and its agencies. Some courts have held that a student's expectation in receiving a high school diploma in return for meeting certain attendance and academic criteria is a form of a property right or liberty interest. *See Debra P. v. Turlington*, 644 F.2d 397 (5th Cir. 1981); *Crump v. Gilmer Indep. Sch. Dist.*, 797 F. Supp. 552, 555-56 (E.D. Tex. 1992); BLACK'S LAW DICTIONARY 500-01 (6th ed. 1990); *see also Procedural due process, Substantive due process. But see Board of Educ. v. Ambach*, 458 N.Y.S.2d 680, (N.Y. App. Div. 1982), *aff'd*, 457 N.E.2d 775 (1983).

Educational necessity—once the party challenging the practice has shown a significant disparate impact, the educational institution using the challenged practice must present sufficient evidence that it is justified by educational necessity. Educational necessity generally refers to a showing that practices or procedures are necessary to meeting an important educational goal. *Elston v. Talladega County Bd. of Educ.*, 997 F.2d 1394, 1412 (11th Cir. 1993) (citing *Georgia State Conf. of Branches of NAACP v. Georgia*, 775 F.2d 1403, 1412, 1417 (11th Cir. 1985)). In the context of testing this means the test or assessment procedure must serve a legitimate educational goal and be valid and reliable for the purpose used.

Equal protection—classifications based on race, sex or other grounds may be challenged under the equal protection clause of the Fourteenth Amendment to the U.S. Constitution when imposed by state or local government agencies. Distinctions explicitly based on race or ethnicity, neutral criteria having a discriminatory purpose, or other intentionally discriminatory conduct based on race or ethnicity will violate the Fourteenth Amendment, unless the action is narrowly tailored to serve a compelling purpose. Intentional sex discrimination will violate the Fourteenth Amendment unless there is an exceedingly persuasive justification. *United States v. Virginia*, 518 U.S. 515 (1996). Distinctions based on other grounds will not violate the equal protection clause unless they are not rationally related to a legitimate governmental objective.

Facially neutral—a regulation, rule, practice or other activity that does not appear to be discriminatory. A facially neutral practice may be found in violation of federal law if the practice results in significant differences in the distribution of benefits or services to persons based on race, national origin, sex or disability without a substantial legitimate educational justification or there are equally or comparably effective alternative practices available that meet the institution's goals with less disparate impact. *See, e.g., Lau v. Nichols*, 414 U.S. 563 (1974); *Larry P. v. Riles*, 793 F.2d 969 (9th Cir. 1984).

High-stakes educational decisions for students—decisions that have significant impact or consequences for individual students. These decisions may involve student placement in gifted and talented programs; decisions concerning whether a student has a disability; the appropriate educational program for a student with a disability; promotion or graduation decisions; and higher education admissions decisions and scholarship awards. National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, pp. 1-2 (Jay P. Heubert & Robert Hauser eds., 1999); *Larry P. v. Riles*, 793 F.2d 969 (9th Cir. 1984); *Sharif v. New York State Educ. Dep't*, 709 F. Supp. 345 (S.D.N.Y 1989).

Less discriminatory alternative—if the education institution presents sufficient evidence that the test use or educational practice in question is justified educationally, the party challenging the test has the opportunity to show that there exists an equally or comparably effective alternative practice that meets the institution's goals and that would eliminate or reduce the adverse impact. *Elston v. Talladega County Bd. of Educ.*, 997 F.2d 1394, 1407 (11th Cir. 1993); *Georgia State Conf. of Branches of NAACP v. Georgia*, 775 F.2d 1403 (11th Cir. 1985). Costs and administrative burdens are among the factors considered in assessing whether the alternative practice is equally effective in fulfilling the institution's goals. *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 661 (1989); *Sharif v. New York State Educ. Dep't*, 709 F. Supp. 345, 363-64 (S.D.N.Y. 1989) (defendant's claim that proposed alternative was not feasible and was excessively burdensome not persuasive since most other states used proposed alternative).

Procedural due process—the right each American citizen has under the Constitution to a fair process in actions that affect an individual's life, liberty or property. Procedural due process includes notice and the right to be heard. Some courts have found that procedural due process applies to the implementation of minimum competency examinations required for high school graduation. *Debra P. v. Turlington*, 474 F. Supp. 244, 263-64 (M.D. Fla. 1979), *aff'd in part and vacated in part*, 644 F.2d 397 (5th Cir. 1981); *Erik V. v. Causby*, 977 F. Supp. 384, 389-90 (E.D.N.C. 1997); *Crump v. Gilmer Indep. Sch. Dist.*, 797 F. Supp. 552, 555-56 (E.D. Tex. 1992); Black's Law Dictionary 1203 (6th ed. 1990).

Significantly disproportionate—when statistical analysis shows that the success rate of members of an identified group is significantly lower than would be expected from random distribution within the appropriate qualified pool, the test in question is said to have a disproportionate adverse impact. There is no set formula to determine when a sufficient level of adverse impact has been reached; the Supreme Court has stated that statistical disparities must be sufficiently substantial that they raise an inference of causation. Courts have advanced percentage disparities, standard deviations or other statistical formulae to address this component. Disparate impact itself does not necessarily mean that discrimination has taken place, but it does trigger an inquiry regarding the educational justification of the challenged practice. *See Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 994-95 (1988); *Groves v. Alabama State Bd. of Educ.*, 776 F. Supp. 1518, 1529-32 (M.D. Ala. 1991); *Richardson v. Lamar County Bd. of Educ.*, 729 F. Supp. 806, 815-16 (M.D. Ala. 1989), *aff'd*, 935 F.2d 1240 (11th Cir. 1991).

Statutory rights—rights protected by statute, as opposed to constitutional rights, which are protected by the Constitution.

Substantive due process—often stated as "fundamental fairness." In an education context, proof that students had not been taught the material on which they were tested might be a substantive due process violation. Some courts have held that students have the equivalent of a property or liberty interest in graduating or being promoted according to the expectations given them. *See Debra P. v. Turlington*, 644 F.2d 397 (5th Cir. 1981); *Crump v. Gilmer Indep. Sch. Dist.*, 797 F. Supp. 552, 555-56 (E.D. Tex. 1992); Black's Law Dictionary 1429 (6th ed. 1990).

Unitary system—a desegregated school system. The Supreme Court has held that all previously intentionally segregated school systems are required to become unitary systems. Although the term has been interpreted in different ways by different courts, a "unitary system" is typically one in which all vestiges of past discrimination and segregated practices have been eliminated. *See Freeman v. Pitts*, 506 U.S. 467, 486-89 (1992); *Board of Educ. v. Dowell*, 498 U.S. 237, 243-46, 249-51 (1991); *Keyes v. School Dist. No. 1*, 413 U.S. 189, 208, 257-58 (1973); *Georgia State Conf. of Branches of NAACP v. Georgia*, 775 F.2d 1403, 1413-16 (11th Cir. 1985); *Bester v. Tuscaloosa City Bd. of Educ.*, 722 F.2d 1514, 1517 (11th Cir. 1984); *Debra P. v. Turlington*, 474 F. Supp. 244, 249-57 (M.D. Fla. 1979) *aff'd in part and vacated in part*, 644 F.2d 397 (5th Cir. 1981).

APPENDIX B: Glossary of Test Measurement Terms

This glossary is provided as a plain language reference to assist readers in understanding commonly used test measurement terms used in this guide or terms relevant to issues discussed in the guide. For additional relevant information, readers are encouraged to review the Glossary in the *Joint Standards*, as well as the appropriate chapters in the *Joint Standards*.

Accommodation—A change in how a test is presented, in how a test is administered, or in how the test taker is allowed to respond. This term generally refers to changes that do not substantially alter what the test measures. The proper use of accommodations does not substantially change academic level or performance criteria. Appropriate accommodations are made in order to level the playing field, i.e., to provide equal opportunity to demonstrate knowledge.

Achievement level/ proficiency levels—Descriptions of a test taker's competency in a particular area of knowledge or skill, usually defined as ordered categories on a continuum, often labeled from "basic" to "advanced," that constitute broad ranges for classifying performance.

Alternate assessment—An assessment designed for those students with disabilities who are unable to participate in general large-scale assessments used by a school district or state, even when accommodations or modifications are provided. The alternate assessment provides a mechanism for students with even the most significant disabilities to be included in the assessment system.

Assessment—Any systematic method of obtaining information from tests or other sources, used to draw inferences about characteristics of people, objects, or programs.

Bias—In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct irrelevant components of test scores. Bias usually favors one group of test takers over another.

Bilingual—The characteristic of being relatively proficient in two languages.

Classification accuracy—The degree to which neither false positive nor false negative categorizations and diagnoses occurs when a test is used to classify an individual or event.

Composite score—A score that combines several scores according to a specified formula.

Content areas—Specified subjects in education, such as language arts, science, mathematics, or history.

Content domain—The set of behaviors, knowledge, skills, abilities, attitudes or other characteristics to be measured by a test, represented in a detailed specification, and often organized into categories by which items are classified.

Content standard—Statements which describe expectations for students in a subject matter at a particular grade or at the completion of a level of schooling.

Content validity—Validity evidence which analyzes the relationship between a test's content and the construct it is intended to measure. Evidence based on test content includes logical and empirical analyses of the relevance and representativeness of the test content to the defined domain of the test and the proposed interpretations of test scores.

Construct—The concept or the characteristic that a test is designed to measure.

Construct equivalence—1. The extent to which the construct measured by one test is essentially the same as the construct measured by another test. 2. The degree to which a construct measured by a test in one cultural or linguistic group is comparable to the construct measured by the same test in a different cultural or linguistic group.

Construct irrelevance—The extent to which test scores are influenced by factors that are irrelevant to the construct that the test is intended to measure. Such extraneous factors distort the meaning of test scores from what is implied in the proposed interpretation.

Constructed response item—An exercise for which examinees must create their own responses or products rather than choose a response from an enumerated set. Short-answer items require a few words or a number as an answer, whereas extended-response items require at least a few sentences.

Construct underrepresentation—The extent to which a test fails to capture important aspects of the construct that the test is intended to measure. In this situation, the meaning of test scores is narrower than the proposed interpretation implies.

Criterion validity—Validity evidence which analyzes the relationship of test scores to variables external to the test. External variables may include criteria that the test is expected to be associated with, as well as relationships to other tests hypothesized to measure the same constructs and tests measuring related constructs. Evidence based on relationships with other variables addresses questions about the degree to which these relationships are consistent with the construct underlying the proposed test interpretations. *See Predictive validity.*

Criterion-referenced—Scores of students referenced to a criterion. For instance, a criterion may be specific, identified knowledge and skills that students are expected to master. Academic content standards in various subject areas are examples of this type of criterion.

Criterion-referenced test—A test that allows its users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others. Examples of criterion-referenced interpretations include comparison to cut scores, interpretations based on expectancy tables, and domain-referenced score interpretations.

Cut score—A specified point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point. *See Performance standard.*

Discriminant validity—Validity evidence based on the relationship between test scores and measures of different constructs.

Error of measurement—The difference between an observed score and the corresponding true score or proficiency. This unintended variation in scores is assumed to be random and unpredictable and impacts the estimate of reliability of a test.

False negative—In classification, diagnosis, or selection, an error in which an individual is assessed or predicted not to meet the criteria for inclusion in a particular group but in truth does (or would) meet these criteria.

False positive—In classification, diagnosis, or selection, an error in which an individual is assessed or predicted to meet the criteria for inclusion in a particular group but in truth does not (or would not) meet these criteria.

Field test—A test administration used to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting. A field test is generally more extensive than a pilot test. *See Pilot test.*

High-stakes decision for students—A decision whose result has important consequences for students.

Internal consistency estimate of reliability—An index of the reliability of test scores derived from the statistical interrelationships of responses among item responses or scores on separate parts of a test.

Inter-rater agreement—The consistency with which two or more judges rate the work or performance of test takers; sometimes referred to as *inter-rater reliability*.

Local evidence—Evidence (usually related to reliability or validity) collected for a specific and particular set of test takers in a single institution, district, or state, or at a specific location.

Local norms—Norms by which test scores are referred to a specific, limited *reference population* of particular interest to the test user (such as institution, district, or state); local norms are not intended as representative of populations beyond that setting.

Norm-referenced—Scores of students compared to a specified reference population.

Norm-referenced test—A test that allows its users to make score interpretations of a test taker's performance in relation to the performance of other people in a specified reference population.

Norms—Statistics or tabular data that summarize the distribution of test performance for one or more specified groups, such as test takers of various ages or grades. The group of examinees represented by the norms is referred to as the reference population. Norm reference populations can be a local population of test takers, e.g. from a school, district or state, or it can represent a larger population, such as test takers from several states or throughout the country.

Percentile rank—Most commonly, the percentage of scores in a specified distribution that fall below the point at which a given score lies. Sometimes the percentage is defined to include scores that fall at the point; sometimes the percentage is defined to include half of the scores at the point.

Performance assessments—Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied.

Performance standard—1. An objective definition of a certain level of performance in some domain in terms of a cut score or a range of scores on the score scale of a test measuring proficiency in that domain. 2. A statement or description of a set of operational tasks exemplifying a level of performance associated with a more general content standard; the statement may be used to guide judgements about the location of a cut score on a score scale. The term often implies a desired level of performance. *See Cut scores.*

Pilot test—A test administered to a representative sample of test takers to try out some aspects of the test or test items, such as instructions, time limits, item response formats, or item response options. *See Field test.*

Portfolio assessments—A systematic collection of educational or work products that have been compiled or accumulated over time, according to a specific set of principles.

Precision of measurement—A general term that refers to a measure's sensitivity to error of measurement.

Predictive validity—Validity evidence that analyzes the relationship of test scores to variables external to the test that the test is expected to predict. Predictive evidence indicates how accurately test data can predict criterion scores that are obtained or outcomes that occur at a later time. *See Criterion evidence of validity; False positive error; False negative error.*

Random error—An unsystematic error; a quantity (often observed indirectly) that appears to have no relationship to any other variable.

Reference population—The population of test takers represented by test norms. The sample on which the test norms are based must permit accurate estimation of the test score distribution for the reference population. The reference population may be defined in terms of size of the population (local or larger), examinee age, grade, or clinical status at time of testing, or other characteristics.

Reliability—The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group.

Sample—A selection of a specified number of entities called sampling units (test takers, items, schools, etc.) from a large specified set of possible entities, called the population. A random sample is a selection according to a random process, with the selection of each entity in no way dependent on the selection of other entities. A stratified random sample is a set of random samples, each of a specified size, from several different sets, which are viewed as strata of the population.

Sampling from a domain—The process of selecting test items to represent a specified universe of performance.

Score—Any specific number resulting from the assessment of an individual; a generic term applied for convenience to such diverse measures as test scores, absence records, course grades, ratings, and so forth.

Scoring rubric—The established criteria, including rules, principles, and illustrations, used in scoring responses to individual items and clusters of items. The term usually refers to the scoring procedures for assessment tasks that do not provide enumerated responses from which test takers make a choice. Scoring rubrics vary in the degree of judgement entailed, in the number of distinct score levels defined, in the latitude given scorers for assigning intermediate or fractional score values, and in other ways.

Selection—A purpose for testing that results in the acceptance or rejection of applicants for a particular educational opportunity.

Sole criterion—When only one standard (such as a test score) is used to make a judgement or a decision. This can include a step-wise decision-making procedure where students must reach or exceed one criterion (such as a cut score of a test) independent of or before other criteria can be considered.

Speed test—A test in which performance is measured primarily or exclusively by the time to perform a specified task, or the number of tasks performed in a given time, such as tests of typing speed and reading speed.

Standards-based assessment—Assessments intended to represent systematically described content and performance standards.

Systematic error—A score component (often observed indirectly), not related to the test performance, that appears to be related to some salient variable or sub-grouping of cases in empirical analyses. This type of error tends to increase or decrease observed scores consistently in members of the subgroup or levels of the salient variable. *See Bias.*

Technical manual—A publication prepared by test authors and publishers to provide technical and psychometric information on a test.

Test—An evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process.

Test developer—The person(s) or agency responsible for the construction of a test and for the documentation regarding its technical quality for an intended purpose.

Test development—The process through which a test is planned, constructed, evaluated and modified, including consideration of content, format, administration, scoring, item properties, scaling, and technical quality for its intended purpose.

Test documents—Publications such as test manuals, technical manuals, user's guides, specimen sets, and directions for test administrators and scorers that provide information for evaluating the appropriateness and technical adequacy of a test for its intended purpose.

Test manual—A publication prepared by test developers and publishers to provide information on test administration, scoring, and interpretation and to provide technical data on test characteristics.

Validation—The process through which the validity of the proposed interpretation of test scores is evaluated.

Validity—The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test.

Validity argument—An explicit scientific justification of the degree to which accumulated evidence and theory supports the proposed interpretation(s) of test scores.

Validity evidence—Systematic documentation that empirically or theoretically demonstrates, under the specific conditions of the individual analysis, to which extent, for whom, and in which situations test score inferences are valid. No single piece of evidence is sufficient to document validity of test scores; rather, aspects of validity evidence must be accumulated to support specific interpretations of scores.

Validity evidence for relevant subgroups—Validity results disaggregated by subgroups, such as by race/ethnicity, or by disability or limited English proficiency status. This type of evidence is appropriate generally when credible research suggests that interpretations of the test scores may differ by subgroup. For instance, if a test will be used to predict future performance, validity evidence should document that the scores are as valid a predictor of the intended performance for one subgroup as for another.

APPENDIX C: Accommodations Used by States

This Appendix lists many of the accommodations used in large-scale testing for limited English proficient students and students with disabilities. The list is not meant to be exhaustive, and its use in this document should not be seen as an endorsement of any specific accommodations. Rather, the Appendix is meant to provide examples of the types of accommodations that are being used with limited English proficient students and students with disabilities.

Table 1
Accommodations for Limited English Proficient Students

PRESENTATION FORMAT

Translation of directions into native language
Translation of test into native language
Bilingual version of test (English and native language)
Further explanation of directions
Plain language editing
Use of word lists/ dictionaries
Bilingual dictionary
Large print

ADMINISTRATION FORMAT

Oral reading in English
Oral reading in native language
Person familiar to students administers test
Clarification of directions
Use of technology
Alone, in study carrel
Separate room
With small group
Extended testing time
More breaks
Extending sessions over multiple days

RESPONSE FORMAT

Allow student to respond in writing in native language
Allow student to orally respond in native language
Allow student to orally respond in English
Use of technology

OTHER

Out-of-level testing
Alternate scoring of writing test

Adapted from: Council of Chief State School Officers, *Annual Survey: State Student Assessment Programs*, Washington D.C., 1999

Table 2
Accommodations for Students with Disabilities

PRESENTATION FORMAT

Braille edition
Large-print editions
Templates to reduce visual field
Short-segment testing booklets
Key words highlighted in directions
Reordering of items
Use of spell checker
Use of word lists/dictionaries
Translated into sign language

ADMINISTRATION FORMAT

Oral reading of questions
Use of magnifying glass
Explanation of directions
Audiotape directions or test items
Repeating of directions
Interpretation of directions
Videotape in American Sign Language
Interpreter signs test in front of classroom/student
Signing of directions
Amplification equipment
Enhanced lighting
Special acoustics
Alone in study carrel

Individual administration
In small groups
At home with appropriate supervision
In special education classes separate room
Off campus
Interpreter with teacher facing student; student in front of classroom
Adaptive furniture
Use place marker
Hearing aids
Student wears noise buffers
Administrator faces student
Specialized table
Auditory trainers
Read questions aloud to self
Colored transparency
Assist student in tracking by placing students finger on item
Typewriter device to screen out sounds
Extended testing time
More breaks
Extending sessions over multiple days
Altered time of day that test is administered

RESPONSE FORMAT

Mark responses in booklet
Use template for recording
Point to response
Lined paper
Use sign language
Use typewriter/computer/ word processor
Use Braille writer
Oral response, use of scribe
Alternative response methods, use of scribe
Answers recorded on audiotape
Administrator checks to ensure that student is placing responses in correct area
Lined paper for large script printing
Communication board

OTHER

Out-of-level testing

Adapted from: Council of Chief State School Officers, *Annual Survey: State Student Assessment Programs*, Washington D.C., 1999

APPENDIX D: Compendium of Federal Statutes and Regulations

This compendium provides a description of the federal nondiscrimination statutes and regulations that are relevant to testing issues and constitute the primary sources of legal authority in the guide. Specifically, this appendix primarily provides information about pertinent federal laws, including Title VI, Title IX, Section 504, Title II of the Americans with Disabilities Act, and the Individuals with Disabilities Education Act.

A. Title VI

Title VI of the Civil Rights Act of 1964, 42 U.S.C. § 2000d, prohibits race and national origin discrimination by recipients of federal financial assistance. For the regulations issued by the Department of Education implementing Title VI, see 34 C.F.R. Part 100. Under the Civil Rights Restoration Act of 1987, OCR has institutionwide jurisdiction over the recipient of federal funds. 42 U.S.C. § 2000d(4) (1989).

The Title VI statute bars only intentionally discriminatory conduct. However, the regulations promulgated under Title VI prohibit the use of neutral criteria having disparate effects unless the criteria are educationally justified and there are no alternative practices available that are equally effective in serving the institution's goals and result in less disparate effects. *See Guardians Ass'n v. Civil Service Comm'n*, 463 U.S. 582 (1983).

The regulations implementing Title VI do not specifically address the use of tests and assessment procedures, but bar discrimination based on race, color or national origin in any service, financial aid or other benefit provided by the recipient. The provision of the Title VI regulation that prohibits criteria or methods of administration that is often applied in testing cases have the effect of discriminating based on race, color, or national origin. 34 C.F.R. § 100.3(b)(2).

See also 34 C.F.R. § 100, Appendix B, Part K (Guidelines for Eliminating Discrimination and Denial of Services on the Basis of Race, Color, National Origin, Sex, and Handicap in Vocational Education Programs) ("if a recipient can demonstrate that criteria [that disproportionately exclude persons of a particular race, color, national origin, sex, or disability] have been validated as essential to participation in a given program and that alternative equally valid criteria that do not have such a disproportionate adverse effect are unavailable, the criteria will be judged nondiscriminatory. Examples of admission criteria that must meet this test or assessment procedure are . . . interest inventories . . . and standardized test or assessment procedures").

93

B. Title IX

Title IX of the Education Amendments of 1972, 20 U.S.C. §§ 1681 *et seq*., prohibits sex discrimination by recipients of federal financial assistance. For the regulations issued by the Department of Education implementing Title IX, see 34 C.F.R. Part 106. As under Title VI, OCR, per the Civil Rights Restoration Act of 1987, has institutionwide jurisdiction over the recipient of federal funds. 42 U.S.C. § 2000d(4) (1989).

In addition to general prohibitions against discrimination, the regulations implementing Title IX specifically prohibit the discriminatory use of test or assessment procedures in admissions, 34 C.F.R. § 106.21, employment, 34 C.F.R. § 106.52, and counseling 34 C.F.R. § 106.36.

See also 34 C.F.R. § 100, Appendix B, part K (Guidelines for Eliminating Discrimination and Denial of Services on the Basis of Race, Color, National Origin, Sex, and Handicap in Vocational Education Programs), discussed above in relation to Title VI.

C. Section 504 of the Rehabilitation Act of 1973

Section 504 prohibits discrimination by recipients of federal financial assistance. OCR enforces Section 504 and its regulations in education programs. The regulations implementing Section 504 contain certain sections that are particularly relevant to testing situations:

34 C.F.R. § 104.4(b)(4) prohibits criteria or methods of administration that have the effect of discriminating against qualified persons with disabilities.

34 C.F.R. § 104.42(b)(2) prohibits admissions procedures by higher educational institutions that make use of any test or criterion for admission that has a disproportionate, adverse impact on qualified individuals with disabilities unless (1) the test or criterion, as used by the institution, has been validated as a predictor of success in the education program or activity and (2) alternate tests or criteria that have a less disproportionate, adverse impact are not shown to be available. 34 C.F.R. § 104.42(b)(3) requires admissions tests used by post-secondary institutions to be selected and administered so as best to ensure that, when a test is administered to an applicant with a disability, the test results accurately reflect the applicant's aptitude or achievement, rather than reflecting the student's disability (except where disability-related skills are the factors the test purports to measure). 34 C.F.R. §§ 104.44(a) and 104.44(d) require higher education institutions to provide adjustments or accommodations and auxiliary aids and services that enable the student to demonstrate the knowledge and skills being tested.

34 C.F.R. § 104.44(a) states that academic requirements that the institution can demonstrate are essential to the program of instruction or to any directly related licensing requirement will not be regarded as discriminatory.

34 C.F.R. § 104.35(b) requires public elementary and secondary education programs to individually evaluate a student before classifying the student as having a disability or placing the student in a special education program; tests used for this purpose must be selected and administered so as best to ensure that the test results accurately reflect the student's aptitude or achievement or other factor being measured rather than reflecting the student's disability, except where those are the factors being measured. These provisions also require that tests and other evaluation materials include those tailored to evaluate the specific areas of educational need and not merely those designed to provide a single intelligence quotient.

D. Title II of the Americans with Disabilities Act (ADA)

Title II of the Americans with Disabilities Act of 1990 (ADA), 42 U.S.C. § 12134, prohibits discrimination on the basis of disability by public entities. Regulations implementing Title II, issued by the U.S. Department of Justice, can be found at 28 C.F.R. Part 35. OCR enforces Title II as to public schools and colleges. Like the Section 504 regulations, the regulations implementing Title II prohibit "criteria and methods of administration which have the effect of discriminating" against qualified persons with disabilities. *See* 28 C.F.R. § 35.130(b)(3). The regulations also require public entities to make reasonable accommodations to policies, procedures, and practices when the modifications are necessary to avoid discrimination unless the public entity can demonstrate that the modification would fundamentally alter the nature of the service, program, or activity. 28 C.F.R. § 35.130(b)(7).

E. Individuals with Disabilities Education Act (IDEA)

The Individuals with Disabilities Education Act (IDEA) contains important provisions related to testing students with disabilities in elementary and secondary schools. IDEA is enforced by the Office of Special Education Programs in the U.S. Department of Education. As amended in 1997, IDEA requires inclusion of students with disabilities in state- and districtwide assessment programs, with appropriate accommodations, if necessary, unless the student's individual education team decides that participation in all or part of the testing program is not appropriate. The student's individualized education program (IEP) should also state any individual modifications in the administration of state- or districtwide assessments of student achievement that are needed in order for the student to participate in such assessment. If the IEP team determines that the student will not participate in a particular state- or districtwide assessment of student achievement (or part of such an assessment), the student's IEP must include statements of why that assessment is not appropriate for the student and how the student will be assessed. IDEA also requires state or local education agencies to develop guidelines for the alternate assessment of the relatively small number of students with disabilities who cannot take part in state- and districtwide tests to participate in alternate assessments. These alternate assessments must be developed and conducted not later than July 1, 2000. 20 U.S.C. §§ 1412(a)(16) and (17), 1413(a)(6), 1414(d)(1)(A) and (d)(6)(A)(ii); 34 C.F.R. §§ 300.138, 300.139, 300.240, 300.347.

APPENDIX E: Resources and References

Office for Civil Rights
U.S. Department of Education

Minority Students and Special Education: Legal Approaches for Investigation (1995).
Provides an overview of the legal theories and approaches employed in OCR investigations examining disproportionate representation of minority students in special education.

Policy Update on Schools' Obligations Toward National Origin Minority Students With Limited-English Proficiency (1991).
Used by OCR staff to determine schools' compliance with their Title VI obligation to provide any alternative language programs necessary to ensure that national-origin-minority students with limited English proficiency have meaningful access to programs. Provides additional guidance for the December 1985 and May 1970 memoranda.

The Office for Civil Rights' Title VI Language-Minority Compliance Procedures (1985).
Focuses on the treatment of limited English proficient students in programs that received funds from the Department.

Identification of Discrimination and Denial of Services on the Basis of National Origin (May 1970) 35 Fed. Reg. 11595.
Clarifies school district responsibilities to limited English proficient students. Memo was the foundation for the U.S. Supreme Court decision *Lau v. Nichols* and was affirmed in that decision.

Office of Elementary and Secondary Education
U.S. Department of Education

Peer Reviewer Guidance for Evaluating Evidence of Final Assessments Under Title 1 of the Elementary and Secondary Education Act (ESEA) (1999).
Informs the states about types of evidence that would be useful in determining the evaluation of assessments under Title 1.

Taking Responsibility for Ending Social Promotion (1999).
Provides strategies for preventing academic failure and gives information about how these strategies can be sustained through ongoing support for improvement.

Handbook for the Development of Performance Standards: Meeting the Requirements of Title 1 (with Council of Chief State School Officers) (1998).
Describes the best practices and current research on the development of academic performance standards for K-12.

Standards, Assessments and Accountability (1997).
Overview of the major provisions under Title 1 of the Elementary and Secondary Education Act.

National Research Council
National Academy Press, Washington D.C.

High Stakes: Testing for Tracking, Promotion and Graduation (Jay P. Heubert & Robert M. Hauser eds., 1999).
Discusses how tests should be planned, designed, implemented, reported and used for a variety of educational policy goals. Focuses on the uses of tests that make high-stake decisions about individuals and on how to ensure appropriate test use.

Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions (Alexandra Beatty, M.R.C. Greenwood & Robert L. Linn eds., 1999).
Four recommendations regarding test use for admission are made to colleges and universities, including a warning to schools to avoid using scores as more precise and accurate measures of college readiness than they are. One recommendation is made to test producers, which is to make clear the limitations of the information that the scores provide.

Testing, Teaching and Learning: A Guide for States and School Districts (Richard F. Elmore & Robert Rothman eds., 1999).
Practical guide to assist states and school districts in developing challenging standards for student performance and assessment as specified by Title I. Discusses standards-based reform and specifies components of an education improvement system, which are standards, assessments, accountability and monitoring the conditions of instruction.

Improving America's Schooling for Language Minority Children: A Research Agenda (Diane August & Kenji Hakuta eds., 1997).
Summary of the schooling and assessment of extensive study of limited English proficient students. Gives state of knowledge review and identifies research agenda for future study. Includes discussion of student assessment and program evaluation.

Educating One and All: Students with Disabilities and Standards-Based Reform (Lorraine M. McDonnell, Margaret J. McLaughlin & and Patricia Morison eds., 1997).
Twelve recommendations are given regarding how to integrate students with disabilities in standards-based reform, including: participation of students with disabilities should be maximized; that any test alterations must be individualized and have a compelling educational justification; include these students' test results in any accountability system; ensure opportunity for students with disabilities to learn the material tested; and use the IEP process for decision-making on the participation of individual students. Recommendations for policy-makers include: revising policies that discourage the inclusion of students with disabilities in high-stake tests; giving parents enough information to make informed choices about participation; monitoring possible unanticipated consequences of participation, both for standardized testing and for students with disabilities; designing realistic standards; and designing a long-term research agenda.

The Use of I.Q. Tests in Special Education Decision-Making and Planning: Summary of Two Workshops (Patricia Morison, S.H. White & Michael J. Feuer eds., 1996).
Report provides a synthesis of the key themes and ideas discussed at workshops, including: an overview of legal, policy and measurement issues in use of I.Q. tests in special education; validity and fairness of I.Q. testing for student classification and placement; alternative assessment methods used in combination with or as substitutes for I.Q. tests.

Responsible Test Use: Case Studies for Assessing Human Behavior (Lorraine D. Hyde, Gary J. Robertson & Samuel E. Krug, et al., eds., 1993).
Casebook for professionals using educational and psychological test data, which was developed to apply principles to proper test interpretation and actual test use. Cases are organized under eight sections: general training, professional responsibility training, test selection, test administration, test scoring and norms, test interpretation, reporting to clients and administrative or organization policy issues.

Test Measurement Standards

American Educational Research Association, American Psychological Association & National Council on Measurement in Education, *Standards of Educational and Psychological Testing* (1999).
Provides criteria for the evaluation of tests, testing practices, and the effects of test use. Begins with discussion of the test development process, which focuses on test developers, and moves to specific test uses and applications, which focus on test users. One chapter centers on test takers.

National Council on Measurement in Education, *Code of Professional Responsibilities in Educational Measurement* (1995).

Association for Measurement and Evaluation in Counseling and Development, *Responsibilities of Users of Standardized Tests* (1992).

Joint Committee on Testing Practices, *Code of Fair Testing Practices in Education* (1988).

Measurement Texts

Educational Measurement (Robert L. Linn, ed., 3rd ed. 1989).
Includes 11 chapters, including Messick's classic chapter on validity, and organizes them in two parts: theory and general principles; and construction, administration and scoring.

Samuel Messick, *Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning*, American Psychologist 50(9) (September 1995).

Gives a new cohesive definition of validity that looks at score meaning and social values. Six perspectives of construct validity are defined: content, substantive, structural, generalizability, external and consequential.

Martha Thurlow, Judy Elliott & Jim Ysseldyke, *Testing Students With Disabilities* (1998). This document provides guidance about how students with disabilities should be included in large-scale tests, considerations about how to select the appropriate accommodations for which students, and discussions about the role of state and local educators in ensuring proper test use, the use of alternate tests, and appropriate reporting considerations.

Rebecca J. Kopriva, Council of Chief State School Officers, *Ensuring Accuracy in Testing for English Language Learners* (2000).
This resource provides guidance to states, districts, and test publishers about developing, selecting, or adapting large-scale, standardized assessments of educational achievement that are appropriate and valid for English language learners. The guide's practical recommendations identify the "who, what, when, why and how" associated with developing, selecting, or adapting tests for institution use, including how to select the appropriate accommodations for which students, how to collect appropriate validity evidence, and a discussion of salient reporting considerations.

Test Publisher Materials

Most test publishers produce materials that explain the appropriate use of their tests. We encourage interested readers to obtain these materials from the publishers of the tests they administer or from publishers of tests in which they are interested. Readers can also contact the Association of Test Publishers, 655 15th St. NW, Washington, D.C., 20005, telephone 202-857-8444 for more information.

Other Resources

There are many books and other materials that might be helpful to educators and policy-makers as they develop policies, and design and implement programs which include the use of tests in making high-stakes decisions for students. The following web sites will provide additional information and links to some of these resources.

Council for Chief State School Officers
http://www.CCSSO.org

The National Center on Education Outcomes
http://www.coled.umn.edu/NCEO

Center for Evaluation, Research, Standards and Student Testing
http://cresst96.cse.ucla.edu

National Clearinghouse for Bilingual Education
http://www.ncbe.gwu.edu